

# Hyperspectral Image Analysis of Food Quality

**Arngren, Morten**

*Publication date:*
2011

*Document Version*
Publisher's PDF, also known as Version of record

# Hyperspectral Image Analysis of Food Quality

Morten Arngren

Kongens Lyngby 2011
IMM-PHD-2011-255

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192, ISBN 978-87-643-0821-1

# Preface

This thesis was prepared at the Department of Informatics and Mathematical Modelling, the Technical University of Denmark (DTU) and FOSS Analytical A/S in Hillerød Denmark in partial fulfilment of the requirements for acquiring the industrial Ph.D. degree in mathematical modelling.

This industrial Ph.D. project was financed by FOSS Analytical A/S and Forsknings- og Innovationsstyrelsen, Ministeriet for Videnskab, Teknologi og Udvikling.

The thesis deals with hyperspectral image analysis applied to food quality assessment. The main focus is to investigate the benefits of using a hyperspectral imaging system over a traditional single spectrum NIR system. This thesis represents a summary of my research with detailed descriptions provided in the associated appendix.

The project was supervised by Assoc. Professor Jan Larsen (DTU), Professor Rasmus Larsen (DTU), and Senior Scientist Per Waaben Hansen (FOSS Analytical A/S).

The thesis consists of a summary report and a collection of four research papers written during the period 2008–2011.

Kgs. Lyngby, May 2011

Morten Arngren

# Abstract

Assessing the quality of food is a vital step in any food processing line to ensure the best food quality and maximum profit for the farmer and food manufacturer. Traditional quality evaluation methods are often destructive and labour intensive procedures relying on wet chemistry or subjective human inspection. Near-infrared spectroscopy can address these issues by offering a fast and objective analysis of the food quality. A natural extension to these single spectrum NIR systems is to include image information such that each pixel holds a NIR spectrum.

This augmented image information offers several extensions to the analysis of food quality. This dissertation is concerned with hyperspectral image analysis used to assess the quality of single grain kernels. The focus is to highlight the benefits and challenges of using hyperspectral imaging for food quality presented in two research directions.

Initially, the visualisation and interpretation of hyperspectral images are discussed. A Bayesian based unmixing method is presented as a novel approach to decompose a hyperspectral image into interpretable components. Secondly, hyperspectral imaging is applied to a dedicated application of predicting the degree of pre-germination in single barley kernels using a customised classification framework. Both contributions serve to illustrate the improvement of adding image information to NIR systems in food quality assessment applications.

# Resumé

Vurdering af kvaliteten af fødevarer er et afgørende skridt i enhver fødevareforarbejdning for at sikre både den bedste kvalitet og maksimal profit for landmanden og fødevare producenten.

Traditionelle metoder til evaluering af kvaliteten er ofte destruktive og arbejdskrævende procedurer, der beror sig på våd kemi eller menneskelig inspektion. Nær-infrarød spektroskopi kan løse disse problemer ved at tilbyde en hurtig og objektiv analyse af fødevarekvaliteten. En naturlig udvidelse af disse enkelt spectrum NIR-systemer er, at inkludere billedinformation, således at hver pixel repræsenterer et NIR spektrum.

Denne ekstra billedinformation kan give flere udvidelser til analysen af fødevarekvalitet. Denne afhandling beskæftiger sig med hyperspektral billedanalyse anvendt til at vurdere kvaliteten af enkelte kornkerner. Dens fokus er, at fremhæve fordelene og udfordringerne ved at bruge hyperspektral billedanalyse til vurdering af fødevarekvalitet og præsenteres i to forskningsretninger.

Indledningsvis diskuteres visualisering og fortolkning af hyperspektrale billeder. En Bayesiansk baseret unmixing metode præsenteres som en ny tilgang til nedbryde hyperspektrale billeder til fortolkelige komponenter. For det andet anvendes hyperspektral billedanalyse til at forudsige graden af forspiring i enkelte bygkerner ved hjælp af en dedikeret klassifikationsmodel. Begge bidrag tjener til at illustrere en forbedring ved at tilføje billedinformation til NIR-systemer til vurdering af fødevarekvalitet.

# Papers included in the thesis

## CONFERENCE PAPERS

[A] Rasmus Larsen, Morten Arngren, Per Waaben Hansen, and Allan Aasbjerg Nielsen. Kernel based subspace projection of near infrared hyperspectral images of maize kernels. *Lecture Notes in Computer Science, vol. 5575/2009, p. 560-569, 2009.* Published.

[B] Morten Arngren, Mikkel N. Schmidt, and Jan Larsen. Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images. *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on, DOI 10.1109/MLSP.2009.5306262, p. 1-6, 2009.* Published.

## JOURNAL PAPERS

[C] Morten Arngren, Mikkel N. Schmidt, and Jan Larsen. Unmixing of Hyperspectral Images using Bayesian Non-negative Matrix Factorization with Volume Prior. *Journal of Signal Processing Systems, vol. 65 (3), p. 479-496, DOI: 10.1007/s11265-010-0533-2, 2010.* Published.

[E] Morten Arngren, Per Waaben Hansen, Birger Eriksen, Jan Larsen, and Rasmus Larsen. Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis. *Journal of Agricultural and Food Chemistry, vol. 59 (21), p. 11385-11394, DOI: 10.1021/jf202122y, 2011.* Published.

## TECHNICAL REPORTS

[D] Morten Arngren, Mikkel N. Schmidt, Jan Larsen.
Supplementary material for Unmixing of Hyperspectral Images using Bayesian Nonnegative Matrix Factorization with Volume Prior.
http://www2.imm.dtu.dk/pubdb/p.php?5836, 2011, Published.

[F] Morten Arngren, Jan Larsen, Per Waaben Hansen, Birger Eriksen, Rasmus Larsen.
Supplementary material for Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis.
http://www2.imm.dtu.dk/pubdb/p.php?5989, 2011, Published.

[G] Morten Arngren. Hyperspectral NIR Camera System.
http://www2.imm.dtu.dk/pubdb/p.php?6026, 2011, Published.

## TOOLBOXES

- MATLAB Toolbox: Nonnegative Matrix Factorization with Volume Constraints, http://www.imm.dtu.dk/pubdb/p.php?5846.
- MATLAB Toolbox: Bayesian Nonnegative Matrix Factorization with Volume Constraints, http://www.imm.dtu.dk/pubdb/p.php?5834.
- MATLAB Toolbox: Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis, http://www.imm.dtu.dk/pubdb/p.php?5990.

# Acknowledgements

The motto of Google's scientific search engine Google Scholar is '*Stand on the shoulders of giants*'. During my Ph.D. I had to stand on the shoulders on many giant individuals to whom I wish to express my appreciation.

My greatest gratitude is extended to my supervisors for their engagement in the project through discussions, guidance and contributions to our papers. Assoc. Professor Jan Larsen (DTU Informatics) for having a deep technical understanding with a sharp critical eye for details in any aspect in my research and for always having a door open even in busy times. Senior Scientist Per Waaben Hansen (FOSS Analytical A/S) for introducing me to the world of NIR spectroscopy and for always assisting me in my research with a practical direction. Professor Rasmus Larsen (DTU Informatics) for guiding me through project and taking the time to make image processing simple for me.

Half of the research were conducted at FOSS Analytical A/S as part of the Industrial Ph.D. My research has not been possible without the help and assistance from several people including my immediate managers Dorthe Kjær Pedersen, Bo Büchmann, Lars Nørgaard and Kim Vejlby Hansen for the financial support and of course everyone in the Team Chemometric Development (TCD) group.

I am further very grateful to Mark Berman from Commonwealth Scientific and Industrial Research Organisation (CSIRO) for introducing me to the spectacular world of hyperspectral image analysis and unmixing back in 2007.

The time I spent at Sejet Planteforædling to conduct my pre-germination experiment was also outstanding. I wish to thank everyone at this facility and

in particular Birger Eriksen, Lars Eriksen and Anette Christensen for their assistance and engagement in the experiment. I further appreciate the help from Dansk Bryggeri Laboratorium to give us single kernel sampling plates for our research.

My colleagues at DTU Informatics also deserves acknowledgement for an inspiring environment. In particular Mikkel N. Schmidt and Morten Mørup for always helping me with details in my research. I truly enjoyed our discussions. Jeppe Frisvad and Carsten Stahlhut for reviewing my thesis and of course the three cake club members Toke, Trine and Bjarne for a 'hyggeligt' office.

My family for supporting me through the project. My late grandmother Lita Lindevard who indirectly gave me the opportunity to study my masters and this Ph.D. degree.

...and most importantly to my lovely girlfriend Vicky, who has taken care our two precious children, Xenia and Mads, while Daddy was in the bunker.

# Abbreviations

| | |
|---|---|
| ANN | Artificial Neural Network. |
| CI | Confidence Interval. |
| IID | Independent and Identically Distributed. |
| IR | Infra Red. |
| kPCA | Kernel Principal Component Analysis. |
| LVMR | Latent Variable Multivariate Regression. |
| MAF | Maximum Autocorrelation Factors. |
| MAP | Maximum a Posteriori. |
| MCMC | Markov Chain Monte Carlo. |
| ML | Maximum Likelihood. |
| MNF | Minimum Noise Fraction. |
| MSC | Multiplicative Scatter Correction. |
| NMF | Non-negative Matrix Factorisation. |
| NIR | Near Infrared. |
| NAPC | Noise Adjusted Principal Component. |
| PCA | Principal Component Analysis. |
| PLS | Partial Least Squares. |
| RTT | Radiative Transfer Theory. |
| SNV | Standard Normal Variate. |

# Contents

Chapter 1

# Introduction

With the ever increasing growth of this planets population the demand for food is correspondingly increasing. The assessment of quality of foods is a vital step in producing foods and becomes even more important to utilize the scarce food resource optimally.

The food quality can be defined at many levels. At first the nutrition of the food must be analysed such as identifying rot/putrid and general contaminations to ensure such degenerated food does not enter any food production. In a second level the screened food can be classified into grades of quality for different uses. Both farmers and food manufacturers are interested in a fair evaluation of the food quality to ensure that the farmer maximises his profit and the food producer can produce food of consistent quality.

Current standard procedures to assess the quality of food are often based on labour intensive, destructive and cost ineffective steps relying on wet chemistry or subjective human inspection, e.g. germination of barley kernels [1, 2, 3].

Automated spectroscopic methods perform quality assessment by analysing spectra from the food sample. These methods include light absorption and scatter spectroscopy [27, 35] and are widely used in the food industry due to the fast and objective analysis approach. Optical spectroscopy in the NIR range has

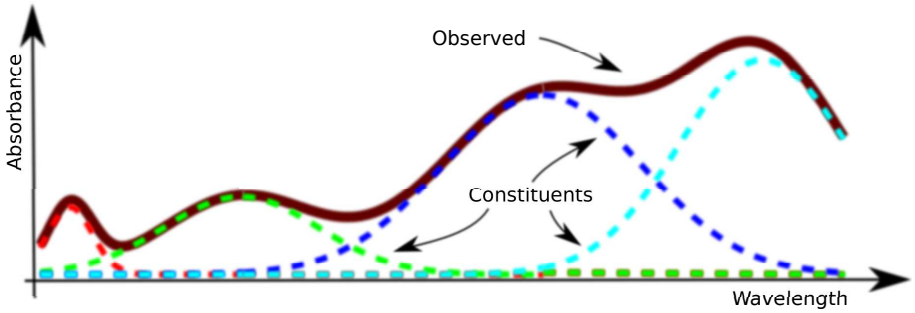



Figure 1.1: Conceptual illustration of observed NIR spectrum made from four underlying spectral signatures of different constituents.

important properties, which makes it particularly useful for food quality assessment. All major food constituents (fat, protein, carbohydrates, and moisture) have strong absorptions in the NIR region. The NIR light can further penetrate the food sample and extract information of the internal constituents of the food (cf. Figure 1.1). In this thesis we will focus on optical spectroscopy in the near infrared (NIR) range based on reflected spectra.

The overall architecture of an automated NIR system to assess food quality is illustrated in Figure 1.2. Initially the food sample is prepared if required and afterwards NIR spectra of the food are acquired. In single spectrum systems this is often an average of several acquisitions. The NIR data is pre-processed prior to entering a mathematical model, which analyses the data and provides interpretation or actual prediction of quality parameters. The mathematical model can be trained on data from a standard quality assessment procedure, e.g. wet chemistry, to provide an automated alternative. The pre-processing of the data is made to enhance any information relevant for predicting the food quality and is often conducted independently of the quality parameter prediction. This step can be incorporated into the mathematical model to improve the interpretation or prediction of the quality parameter. The performance of such a system is never better than the weakest step. Preparing the food and acquiring the NIR spectra are more or less always conducted according to the same procedure for a specific type of food. The pre-processing and mathematical model are two steps, where the systems performance can be modified for improvement. The research contribution of this thesis hence lies primarily in mathematical modelling (marked boldface in Figure 1.2).

The majority of food samples consist of large regions with different constituents. For instance, a meat cut has fat, muscle, and connective tissue regions of very irregular shapes or fungal infections on the surface of grain kernels only reside

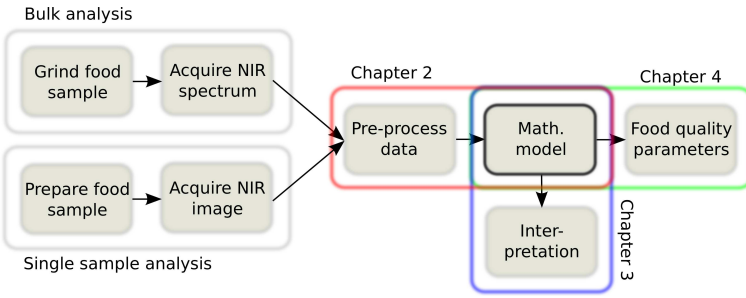



Figure 1.2: Flowchart of the typical pipeline of steps in an automated NIR system for analysing bulk and single food samples.

at small areas. Hence, it can be difficult to obtain a representative sample of the food for a single spectrum NIR analysis, which usually only analyses a small volume of the sample (a few milliliters or less). Food samples are therefore ground down to a small particle size before the NIR analysis is conducted. This complicates the analysis and eliminates the structure of the sample, which may contain important information. In addition, any constituents in low concentrations are inappropriately diluted as a consequence and may be lost in the observed NIR spectrum. Single spectrum NIR analysis is furthermore often based on a relatively large volume of samples, e.g. in grain applications a bulk analysis is conducted and not on single kernels.

Physical sampling constitutes a major problem in NIR analysis and other spectroscopic methods [86]. Digital imaging addresses the issue by providing spectral information at each individual point/pixel of an image (cf. Figure 1.3) without requiring physical pretreatment of the sample. A natural extension to spectrum based analysis is hence to incorporate image data to combine spectral and spatial information. This preserves the spatial structure, while capturing detailed spectral information from the sample and has the advantage of being non-destructive compared to single spectrum NIR analysis.

These spectral images often have more wavelength bands than the classic three bands used in photography (red, green and blue) and are denoted *multispectral* or *hyperspectral* depending on the number of bands. Figure 1.3 shows a conceptual illustration of a NIR image of barley kernels with $M$ bands. For small $M$ the spectra are discrete and the image is referred to as multispectral and for large $M$ the spectra are considered continuous and is denoted hyperspectral[1]. The notions of multi- and hyperspectral are in practice often divided by the technology behind the image acquisition. One approach, where a 2D image is

[1]Strictly speaking spectra in hyperspectral images are also discrete, but represented by numerous bands.

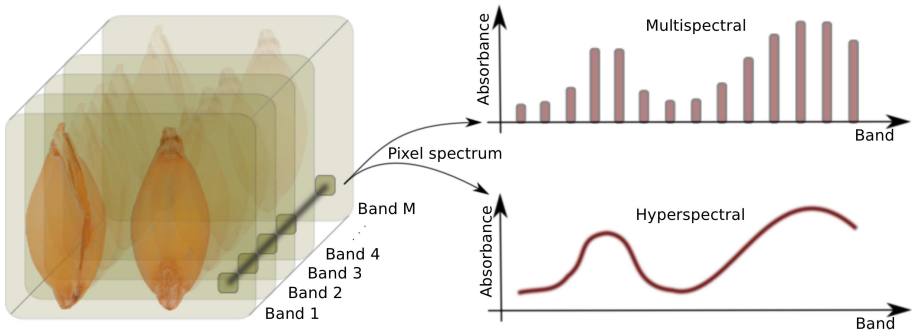



Figure 1.3: Conceptual illustration of a multi- and hyperspectral image with $M$ bands. Each band is represented by a monochrome image and each pixel holds a spectrum.

acquired at each wavelength produces multispectral images at discrete bands. Images can also be acquired by passing the light through an optical grating and projecting it onto an image sensor (pushbroom principle) to produce hyperspectral image with continuous bands. The immediate advantage of multispectral system is the lower number of bands to acquire meaning a faster and most likely cheaper system. The acquisition system can be tuned to specific wavelengths relevant for the application at hand. Due to quantum mechanical properties NIR spectra can be characterised by a soft and smooth signature compared to other IR ranges (this will be elaborated in Chapter 2). This means that only few bands may be required to obtain the necessary information for the task at hand and hence this property justifies multispectral systems. However, spectra with only few bands lack the interpretation offered in continuous spectra from hyperspectral images and are more sensitive to measurement noise. Food constituents may have overlapping spectral features at several bands (e.g. protein and starch). A quality parameter prediction model using hyperspectral data is hence more robust towards spectral variance and noise compared to discrete spectra with few bands in such cases. For research purposes hyperspectral imaging systems are ideal to investigate the significance of each wavelength for food applications. Practical NIR instruments for food analysis are often multispectral with few dedicated wavelengths to reduce cost. Table 1.1 list a comparison of the properties of multi- and hyperspectral image systems.

The supplementary image information provides a set of benefits and opportunities compared to single spectrum NIR systems. Initially spectral image analysis offers a visual interpretation of the food sample, which can support the analysis or serve in a direct quality assessment. In addition smaller food samples can be analysed with higher detection sensitivity to capture low concentration constituents. For instance any surface contamination on a single grain kernel

| Property | Multispectral | Hyperspectral |
|---|---|---|
| No. bands | + | +++ |
| Instrument cost | $ | $$$ |
| Acquisition speed | +++ | + |
| Flexibility | + | +++ |
| Spectral interpretation | + | +++ |
| Data load | + | +++ |
| Measurement noise | Sensitive | Robust |

Table 1.1: Comparison of properties of multi-and hyperspectral image systems.

spanning a few pixels can easily be detected contrary to an average spectrum from a ground food sample. The data size of hyperspectral images is naturally much larger than single spectrum NIR. Any analysis will therefore require additional processing of the data leading to longer analysis time. In assessing the quality of single grain kernels a major challenge is to obtain corresponding references for each kernel. This is often difficult since standard reference methods require a minimum amount of substance to provide a reliable result.

Extending single spectrum NIR systems to include image information is an obvious choice in food quality assessment. The overall advantages of hyperspectral image analysis are clear, but also lead to challenges, which must be considered. The ideal mathematical modelling approach is to utilise the added spatial information to achieve improved performance in both prediction of quality parameters and interpretation of the observation.

This thesis will focus on hyperspectral NIR image analysis for assessing the quality of grain kernels and analyse how it can be utilised in such applications. The research will be investigated by considering the following points.

- **Benefits and challenges.**
  Identify benefits and challenges of adding image information to single spectrum NIR systems.

- **Interpretation of hyperspectral images of foods.**
  Food analysis and interpretation of hyperspectral images in terms of constituent concentrations are not new disciplines, whereas the combination has not been addressed before. A framework for interpreting hyperspectral images of foods in terms of constituent concentrations is investigated.

- **Pre-germinated barley application.**
  Illustrate the potential of hyperspectral image analysis in the food quality

application of predicting the degree of pre-germination in single barley kernels.

In practical terms the outcome of the research can be used to both interpret hyperspectral images and enable optimal predictive decision support in relation to single grain kernel analysis. Commercially the research provides a set of general tools for analysis of hyperspectral images and can be used to highlight the general challenges in future development of innovative and cost-effective food instruments. It is not the scope to describe foods in details other than in general terms relevant for the application at hand.

The thesis is presented with the following structure

- **Chapter 2** introduces hyperspectral image analysis and relates to the technical report in Appendix G. The basics of near-infrared spectroscopy and light interaction are described, which leads to a simple interpretation approach of the hyperspectral images. The NIR hyperspectral line-scan camera system is further introduced and a few relevant pre-processing steps are presented.

- **Chapter 3** discusses visualisation tools to identify food constituents and relates to the papers in Appendix A, B, C, and D. A novel method is presented for decomposing hyperspectral images into underlying internal food constituents, such that the concentration of constituents can be visualised and quantified. The practical use of the model is demonstrated by analysing single wheat kernels. Our approach is based on the spectral information in the images and is interpreted spatially. In terms of food quality our model does not provide a direct quality measure, but can accommodate a visualisation of the internal structure or aid in analysing the food quality. It can serve as an analysing tool in conjunction with a food quality assessment application.
- **Chapter 4** presents a food quality assessment application using hyperspectral imaging and relates to the papers in Appendix E and F. A modelling framework is described to predict the degree of pre-germination of single barley kernels with the focus on optimal performance and not necessarily interpretation of features. The information in single spectrum NIR systems is also present in hyperspectral NIR images as the spatial average. We will not make quantitative comparisons between these two systems, but merely emphasize the potential of hyperspectral NIR imaging in food quality applications.
- **Chapter 5** concludes the research with a summary.

Chapter 2

# Hyperspectral Image Analysis

Classic image acquisition and analysis is based on three color bands (red, green, and blue), which is sufficient for human visualization. In the context of identifying or extracting material constituents of eg. foods, three wide channels are rarely enough. Hyperspectral image analysis can include more than 50 channels and hence provides the opportunity to capture more detailed spectral information required in analysis of foods. Figure 1.3 shows a conceptual illustration of a hyperspectral image, where each band is represented by a monochrome image and each pixel holds a spectrum.

Hyperspectral image analysis has been widely used in remote sensing applications, eg. [28, 44, 57, 100, 106]. These applications have typically been limited to organisations with sufficient funds such as NASA. Recently numerous data sets have been made publicly available for analysis based on satellite image data, eg. AVIRIS[1] [44, 106] (cf. Plaza *et al.* for a more comprehensive list of data sets [88]). Advances made in sensor technology has lead to more cost-effective and commercial hyperspectral instruments being manufactured by companies, such as Headwall Photonics Inc.[2] (MA, US), SPECIM[3] (Finland), and NEO[4] (Norway). The increase of computational power in recent years offers more

[1] Free data available for download at http://aviris.jpl.nasa.gov/html/aviris.freedata.html
[2] http://www.headwallphotonics.com
[3] http://www.specim.fi/
[4] http://www.neo.no/


Figure 2.1: The electromagnetic spectrum of the infrared range (IR) above the visual region (cf. ISO20473). Other IR range divisions exist, eg. CIE and DIN.

complex analysis tools and methods to be developed and used in analysing hyperspectral images. The field of hyperspectral image analysis has therefore received wide attention and has contributed with research in many areas. New analysis methods have been developed and provide tools for advanced analysis, such as extracting constituents and their concentrations from the images [7, 14, 18, 28, 78, 84, 88, 110]. The tools are applied in many application areas and include remote sensing [28, 57, 102, 105, 106], mining [12] and food applications, such as analysing wheat kernels [13, 29, 42, 64, 99, 101].

In the remainder of this chapter we will introduce basic NIR spectroscopy and simple light interactions, which forms the basis behind hyperspectral image analysis. The linear mixing model will be presented along with a visual approach to interpret the model. Finally the hyperspectral camera system will be introduced and relevant pre-processing steps will be shortly described.

## 2.1 Near Infrared Spectroscopy

The field of spectroscopy concerns the interaction between radiated energy and matter and can be exploited to analyse the constituents in materials. The interaction can be described as vibrations in molecules excited within the electronic ground state of the molecule to a vibration state by influence of external induced electromagnetic energy.

The molecular vibrations are sensitive to electromagnetic energy in the IR range, which lies above microwaves and below the visual range in the electromagnetic spectrum, cf. Figure 2.1. The IR range covers a wide range from the edge of the red visible light at 780nm to 1000$\mu$m. The NIR range lies from 780nm to 3000nm and the mid-IR range is located above 3000nm, where most of the vibrational signatures of objects lie.

The mid-IR vibrations can be divided in two groups. The low frequency vibrations affected across the molecule (fingerprint region [27, 104]) and the vibrations from specific functional groups part of the molecule. The latter represent the

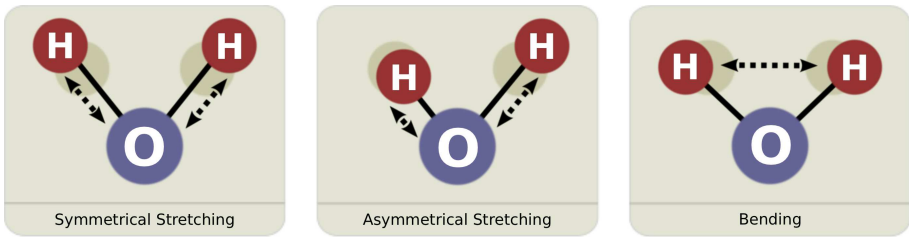



Figure 2.2: The three different vibrational modes of a $H_2O$ molecule. The gray $H$ atom in the background represent the stationary state.

strongest vibrations and exist in many configurations [27, 104]. A general non-linear structured molecule has $3N-6$ vibration modes, where $N$ is the number of atoms. A common molecule in food chemistry is water, which has three vibrational modes as illustrated in Figure 2.2. Other molecules may have additional vibrational modes depending on their molecular structure, eg. the $CH_2$ group has six $C-H$ vibration modes.

The fundamental tones of the molecular vibrations of functional groups lie in the IR range from 2500nm to approx. 7700nm and are characterised by narrow peaks at distinct wavelengths depending on the material and molecular structure. The harmonic overtones and combination bands ripple down the electromagnetic spectrum and appear in the NIR range as smoother spectral observations. They exhibit a high degree of overlap in the NIR region and the spectral signatures of specific materials become smoother and hence more difficult to identify. In addition the spectral absorption features often appear in the same NIR region, which makes separation and interpretation difficult.

One of the important properties of NIR light is the larger penetration depth into a sample compared to the visual range, where surface reflections are more dominant [104]. This means that NIR light has the ability to penetrate an object and extract information of the internal compounds. The penetration depth is highly dependent on the specific wavelength, intensity, and the interacting material.

In practice a hyperspectral image system includes a sensor to detect the NIR reflected light. Various sensors exist based on different sensitive materials and operate in different spectral ranges, eg. the cost effective InGaAs based sensors are sensitive from approx. $900-1700$nm[5] and PbS-based sensors are sensitive from approx $1000-2500$nm[6].

[5]Information on InGaAs sensors are available at manufacturer http://www.hamamatsu.com
[6]Information on PbS sensors are available at manufacturer http://www.calsensors.com

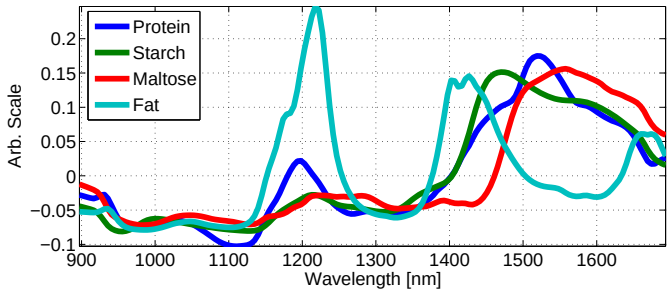



Figure 2.3: NIR spectra of common food constituents showing smooth curvature due to the Mid-IR overtones and combination bands.

### 2.1.1 Food Applications

In food applications the IR technology is widely used since the multi-component analysis approach is fast, non-destructive and cost-effective. In an IR system the macroscopic constituents are detected, whereas microscopic compounds can be too low a concentration for detection. The common chemical constituents detected in food include *Water*, *Protein*, *Carbohydrates*, and *Fat*. The spectra of these basis compounds are shown in Figure 2.3 and are acquired by the hyperspectral camera system used in the research (cf. Appendix B, C and G).

Water exhibits a very strong absorption in the NIR range and is often difficult to measure, as the reflected light can be weaker than the sensor sensitivity. Water can therefore dominate the measured spectrum and the subsequent analysis. It is hence important to reduce the influence of water in food analysis applications, unless water is the analyte to analyse. This can either be done by designing the experiment accordingly to reduce correlation to variables of interest or by post-processing elimination.

In a traditional IR system with a single sensor, the food is often ground to achieve an average spectrum. This means constituents in low concentrations will be diluted, eg. fungal infection on cereals. Imaging systems has the advantage of being able to detect small concentrations spatially separated.

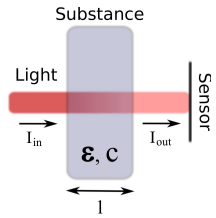



Figure 2.4: A light beam is transmitted through a substance and the observed intensity is attenuated according to Lambert-Beer's law.

## 2.2 Light Interaction

The interaction between light and matter can be reviewed to describe the mixing process and the information contained in the observed data. In this section we will give a small overview of the basic interaction models for both light transmittance and reflectance.

### 2.2.1 Light Transmission

One of the simplest models is light transmission through a substance, where the interaction can be described by Lambert-Beer's law (cf. Figure 2.4). If we transmit light $I_{in}$ through a substance and observe the intensity $I_{out}$ Lambert-Beer's law states

$$I_{out} = I_{in} \exp(-\varepsilon l c) \quad \Leftrightarrow \quad A = -\log\left(\frac{I_{out}}{I_{in}}\right) = -\log(T) = \varepsilon l c, \tag{2.1}$$

where $\varepsilon$ designates the molar absorption coefficient of the substance, $l$ denotes the light path length and $c$ represents the concentration. The transformation from transmittance $T$ to absorbance $A$ by $A = -log(T)$ is conducted to achieve linearity between the specific material concentration $\alpha$ and the observation $A$. The law is associated with a set of conditions. A few of these assumptions dictate the substance must be homogeneous, that no light scattering is present, and that the length of the light path is the same for all incident photons. For homogeneous liquids and gasses these conditions are not hard to comply with. For solid materials light penetration may not be possible and such materials are analysed in a reflection system.

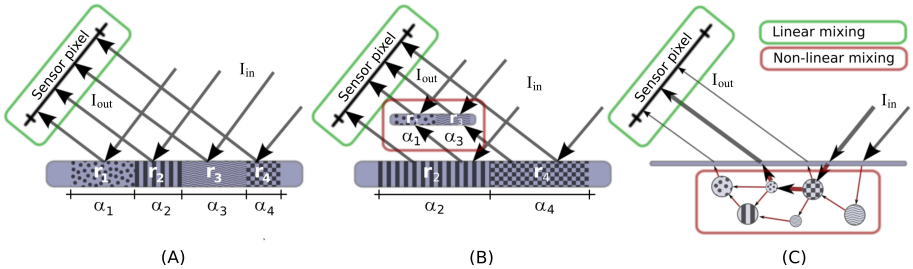



Figure 2.5: Conceptual illustrations of light reflection on object surfaces (A-B) and subsurface scattering (C). (A) The surface is made from different constituents spatially separated and the incident light only has one interaction (surface reflection). This leads to linear mixing of the constituents by the spatial integration on the sensor. (B) The scene has canopy materials, where some incident light have interactions with several materials (multilayer scene), which contribute with non-linear mixing. (C) The incident light penetrates the surface and interacts with several constituents before detection by the sensor giving rise to non-linear intimate mixing.

### 2.2.2 Light Reflection

In a reflecting system the conditions of Lambert-Beer's law can be difficult to fulfil. The length of the light path is not known and the material may not be homogeneous. This gives rise to a different description of the light interactions. Depending on the geometry of the scene and the mixing scale, the observed mixture can be either linear or non-linear. The simplest approach is to assume incident photons only interact with one surface material before detection by the sensor (checkerboard type scenes) and to disregard any material penetration as illustrated in Figure 2.5(A). When the mixing scale in this case is macroscopic and the interaction between the spectral signatures is negligible the mixing becomes linear [4, 16, 46, 55, 77, 98]. If we consider a surface made from $p$ spatially segregated patterns with spectral signature vectors $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_p\}$, where $\boldsymbol{r}_i \in \mathcal{R}^M$, the observed intensity $I_{out}$ can be expressed as

$$\begin{aligned} I_{out} &= I_{in} \cdot (\alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 + \ldots + \alpha_p \boldsymbol{r}_p) \\ \frac{I_{out}}{I_{in}} &= \sum_{i=1}^{p} \alpha_i \boldsymbol{r}_i, \end{aligned} \tag{2.2}$$

where $\alpha_i \in \mathcal{R}$ denote the proportion for each material $i$ corresponding to their relative spatial size. In this case any subsurface scattering is predominated by surface reflection components and hence the observed spectrum is a linear combination of the materials spectral signatures weighted by their fractional

area coverage [4, 16, 46, 55, 77, 98]. The condition of surface reflections is assumed predominant in remote sensing applications and hence the linear mixing in the reflection domain is widely used in these applications, eg. [4, 77, 93].

When the incident light faces multiple interactions with different materials then non-linear mixing occur [16, 19, 91]. This can be illustrated in the two scenarios illustrated in Figure 2.5(B-C).

In the first example the scene includes reflecting canopy materials in two layers in Figure 2.5(B), which causes multiple interactions among the scatterers. The two layers or second order scattering in this case can be expressed by extending Eq. (2.2)

$$\begin{aligned}\frac{I_{out}}{I_{in}} &= \alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 + \ldots + \alpha_1\alpha_2 \boldsymbol{r}_1 \boldsymbol{r}_2 + \alpha_2\alpha_4 \boldsymbol{r}_2 \boldsymbol{r}_4 + \ldots \\ \frac{I_{out}}{I_{in}} &= \sum_{i=1}^{p} \alpha_i \boldsymbol{r}_i + \sum_{i,j=1, i\neq j}^{p} \alpha_i \alpha_j \boldsymbol{r}_i \boldsymbol{r}_j. \end{aligned} \tag{2.3}$$

The second order interactions contribute with non-linear terms and make the mixing more complex. In remote sensing applications of satellite images canopy materials, such as trees, are very common. The model can easily be extended to include higher order scattering non-linear components. However, they are often assumed negligible [77] due to the more complex model and since they typically have insignificant contribution depending on the application. A few attempts have been made to include the multiple scattering, mainly as second order scattering [19, 26, 77, 91, 103, 111].

The second example in Figure 2.5(C) illustrates a scene, where the incident light penetrate the surface and interacts with several materials in close proximity. This leads to intimate mixtures caused by multiple interactions, scatter effects, and varying particle sizes [16, 55]. The non-linear model for these intimate mixtures is much more complex than the linear approximation and is addressed by The Radiative Transfer Theory (RTT), which is a well established macroscopic description and mathematical model for the photon interactions with the materials in the scene [24]. The model depends on scene parameters often hard to obtain and any use typically relies on model approximations. Some of the early work includes Kubelka and Munk, who address the combined effects of scattering and absorption [56] and introduced a model for diffuse reflection. This is widely used to interpret diffuse reflection spectra in chemistry [74]. Hapke derives an approximate analytical solution to the RTT and presents a bidirectional reflectance model [45], which has formed the basis for subsequent research, eg. [53, 62, 90].

Despite the simplicity of the linear model, it has become an acceptable ap-

proximation of light interactions mechanisms, mainly since it has proven very effective in hyperspectral imaging applications and since the non-linear model often becomes too complex, eg. [8, 16, 54, 55, 66, 85].

Most of the hyperspectral imaging applications employing the linear mixing model is analysing remote sensing data, such as satellite images, eg. AVIRIS [44, 106]. In a laboratory environment the distance from sensor to target is much shorter and hence the pixel size is much smaller. This means that the spatial integration in the sensor leading to the linear mixing may not be predominant. For such applications the intimate mixing in a single pixel can be approximated by extending Eq. (2.3) to include multiple scattering non-linear terms. For three interactions this can be expressed as

$$\begin{aligned}\frac{I_{out}}{I_{in}} &= \alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 + \alpha_3 \boldsymbol{r}_3 + \alpha_{12} \boldsymbol{r}_1 \boldsymbol{r}_2 + \alpha_{123} \boldsymbol{r}_1 \boldsymbol{r}_2 \boldsymbol{r}_3 + \dots \\ &= \sum_i \alpha_i \cdot \boldsymbol{r}_1^{\beta_{i,1}} \boldsymbol{r}_2^{\beta_{i,2}} \boldsymbol{r}_3^{\beta_{i,3}}, \end{aligned} \tag{2.4}$$

where $\beta_{i,j}$ are integer values and denote the proportion of each material in the non-linear mix. The model can be simplified by considering the attenuation of light in an optical system, which are caused by both absorption by the material and scattering effects [52, 109]. This is expressed in the attenuation coefficient $\sigma_{att}$ given by

$$\sigma_{att} = \sqrt{\frac{\sigma_{abs}}{D}}, \tag{2.5}$$

where $\sigma_{abs}$ is the absorption coefficient and $D$ is the diffusion coefficient. Initially we assume the bulk scattering properties within a pixel are homogeneous due to the smaller pixel size, ie. $\sigma_{abs}$ and $D$ are constant. This implies that only one light path is predominant in the observed intensity (cf. Figure 2.5(C)) and means the linear integration in the sensor can be neglected. Hence only one term in Eq. (2.4), denoted $\hat{i}$, is significant and $\alpha_{\hat{i}} \approx 1$. We further assume the diffusion coefficient $D = \frac{1}{\sigma_{abs}}$ such that $\sigma_{att} = \sigma_{abs}$, ie. the attenuation of light is solely due to absorption by the material and not subsurface scattering effects. Otherwise we are unable to infer constituent concentrations from the observed data. These assumptions are similar to the light transmission mechanism involving Lambert-Beer in Eq. (2.1) and thus transforming the data to absorbance for linearisation seems obvious. Under these assumptions the mixing reduces to

$$\begin{aligned}\log\left(\frac{I_{out}}{I_{in}}\right) &= \hat{\beta}_1 \log(\boldsymbol{r}_1) + \hat{\beta}_2 \log(\boldsymbol{r}_2) + \hat{\beta}_3 \log(\boldsymbol{r}_3) + \dots \\ &= \sum_n \hat{\beta}_n \log(\boldsymbol{r}_n), \end{aligned} \tag{2.6}$$

where $\hat{\beta}$ denote the fractional abundance, ie. $\hat{\beta}_n = \frac{\beta_n}{\sum_{\bar{n}} \beta_{\bar{n}}}$. In hyperspectral imaging applications in laboratory environment it is clear that non-linear processes are more dominant than for remote sensing applications, due to the mixing scale. The question of whether a set of observed spectra is dominated by linear or non-linear mixing process remains an unresolved issue [55, 100]. In practice the best procedure is an empirical approach, where both the linear and non-linear unmixing models are evaluated. The decomposition can then be compared to a reference set of both spectral components and fractional abundances for best performance.

If the non-linear mixing is considered insignificant and the linear model is in focus, we have presented two modelling approaches each with a set of assumptions. The first operates in the reflectance domain and assumes surface reflections are predominant with no subsurface scattering effects. The second transforms the observed data into the absorbance domain under the assumption of surface penetration with only one significant non-linear mixing light path. In food applications such as ours the latter is more appropriate and is used henceforth.

## 2.3 Linear Mixing Model

Under the assumption of linear mixing of the constituents, the linear mixing model can be formed. Given an acquired hyperspectral image with $M$ bands as $N$ observed non-negative spectra or pixels as the $M$ dimensional columns in the matrix $\boldsymbol{X} \in \mathcal{R}^{M \times N}$, the linear mixing can then be redefined from Eq. (2.6) as a rank $K$ linear matrix factorization,

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{\epsilon}, \tag{2.7}$$

where $\boldsymbol{W} \in \mathcal{R}^{M \times K}$ represent the pure spectral signatures or endmembers as columns (corresponds to $\hat{\beta}_n$ in Eq. (2.6)), matrix $\boldsymbol{H} \in \mathcal{R}^{K \times N}$ denote the concentration or fractional abundances for each pixel as column vectors (corresponds to $\log(\boldsymbol{r}_n)$ in Eq. (2.6)), and $\boldsymbol{\epsilon}$ is white Gaussian noise matrix assumed independent and identically distributed (i.i.d.) Each of the observed pixels $\boldsymbol{x}_n$ can thus be considered a mix of latent pure constituents (cf. Figure 1.1).

In addition to the non-negativity constraint on the observed spectra $\boldsymbol{X}$, the fractional abundances in $\boldsymbol{H}$ must further lie between 0 and 1 and sum to one[7] in order to maintain a proper interpretation. The constraints imposed on the

[7] In the literature, the sum to one constraints on the fractional abundances is sometimes referred to as an *additivity constraint*.

matrix factorization can therefore be expressed as

$$h_{kn} \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} h_{kn} = 1. \tag{2.8}$$

A non-negativity constraint can also be applied to the endmembers as $w_{mk} \geq 0$, since negative absorbance spectra implies the material is emitting light. This corresponds to adding a common offset of all spectra $\boldsymbol{X}$ and does not conflict with interpretation of any spectral features.

## 2.4 Simplex structure

The linear model and the assumptions can be visualised by using the vector space model. The assumptions of non-negativity and sum-to-one of the fractional abundances implies the columns of $\boldsymbol{H}$ lie on a multidimensional simplex defined for $K$ components as (cf. Figure 2.6 for $K = 3$)

$$\left\{ \boldsymbol{h} \in \mathbb{R}^K \middle| \sum_{k=1}^{K} h_k = 1, h_k \geq 0 \right\} \tag{2.9}$$

The linear mixing and constraints further lead to the modelled data lying on a simplex in the observation space, where the pure spectral signatures are the vertices of the simplex and are hence denoted the endmembers. The simplex structure of the observed data is illustrated for three endmembers in Figure 2.7. The mixing model transforms the observations in $\boldsymbol{X}$ to the fractional abundances $\boldsymbol{H}$ as points confined to the unit simplex. Any observation outside violate the constraints and will be forced to the periphery of the simplex to obey the constraints and will hence induce an error in explaining $\boldsymbol{X}$. The sum-to-one constraint further reduces the degrees of freedom of the fractional abundances by one, as three endmembers will lie on a two dimensional simplex, cf. Figure 2.7.

Ideally the observed spectra will span the entire unit simplex and have a full mixing profile. This is however rarely the case in real world observed data, as certain constituents might always mix or in contrast appear in pure form. This is often evident in remote sensing applications of satellite data, where some materials always appear pure with no mixing of other constituents [28, 57]. In food applications the lack of mixing often occurs among the constituents, eg. oil and water. In contrast some constituents are perhaps never observed as pure compounds as they always mix with other constituents. In maize kernels the starch and protein lie in a matrix structure and is not found in pure form. If two or more constituents are positively correlated their spectral signatures collapse

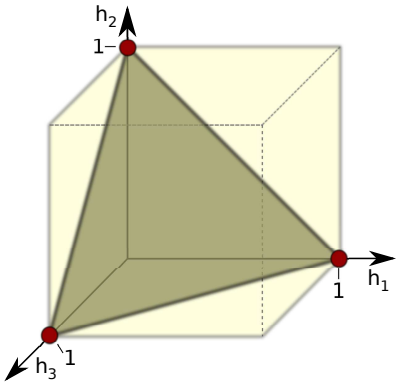



Figure 2.6: Illustration of the unit two-simplex in $\mathbb{R}^3$. The abundance vectors (columns of $\boldsymbol{H}$) are confined to the unit simplex due to the non-negativity and sum-to-one constraints.

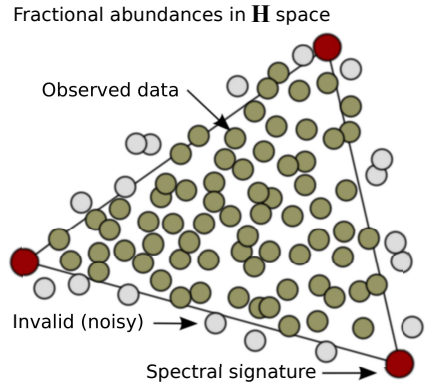



Figure 2.7: A two-simplex in $\boldsymbol{H}$ space with three endmembers (marked red). The data points lying outside the simplex violate the constraints and will induce an error.

into a single endmember and do not span any space inside the simplex and hence become impossible to separate in an unsupervised decomposition. These non-full mixing profiles can be illustrated as in Figure 2.8, where the different mixing profiles have left areas with no observations. Depending on the mixing profile the estimated endmembers and fractional abundances may not be the true set. This becomes particular evident when the mixing level is high (cf. Figure 2.8 profile B), since the true endmembers are far from the observed data. Unsupervised unmixing algorithms including those presented in this thesis are sensitive to the mixing profile of the observed data, eg. [50, 61]. An obvious approach to extract the true endmembers in such case is to include a spectral library to designate the approximate positions of the endmembers. These pure spectral signatures in a library are, however, often expensive and difficult to obtain. A more flexible approach is to include prior information as a library including rough wavelengths of the individual constituents. The information can be based on prior spectroscopic or biological knowledge and not necessarily stem from acquisitions of constituents in pure form. Such a library can act as a weak spectral guide and can be built faster and easier than a precise endmember library.

A major discipline in hyperspectral image analysis is the decomposition of observed data $\boldsymbol{X}$ into the underlying true spectral signatures or endmembers $\boldsymbol{W}$ used to generate the dataset and associated fractional abundances $\boldsymbol{H}$, referred to as *unmixing*. Unmixing is a blind source separation problem, since both the

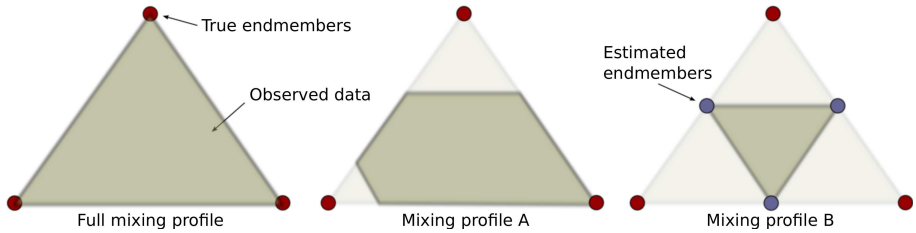



Figure 2.8: The mixing profile of the observed data may not always be full as profile A and B shows. If the mixing profile is not entirely full (profile A) the true fractional abundances can be estimated. If the observed data is highly mixed (profile B), then the true endmembers and fractional abundances can be difficult to extract.

endmembers and fractional abundances are unknown and must be inferred from the observations. The field of unmixing will be introduced in details in Chapter 3 in conjunction with our unmixing method contribution.

## 2.5 Hyperspectral Camera System

Our research utilises a hyperspectral camera system from Headwall Photonics Inc. (MA, US) to acquire the hyperspectral images as the data sets. The heart of the system is a hyperspectral line scan NIR camera produced by XenICs (Leuven, Belgium) based on an InGaAs sensor sensitive from $900 - 1700$nm in 165 bands. It further includes a conveyor belt from miCos GmbH (Eschbach, Germany) for moving the sample past the camera scan line, a NIR light source from TechniQuip (CA, US), and a framework for fixating the system. The acquisition procedure is based on the push broom principle, where the sample moves past the camera view slot and a scan line is acquired. The scanned lines are afterwards assembled to comprise the acquired hyperspectral image.

The hyperspectral NIR camera system is described in details in the technical report (cf. Appendix G). The general properties of the system are presented in terms of spatial and spectral performance along with a description of the standard operating procedure.

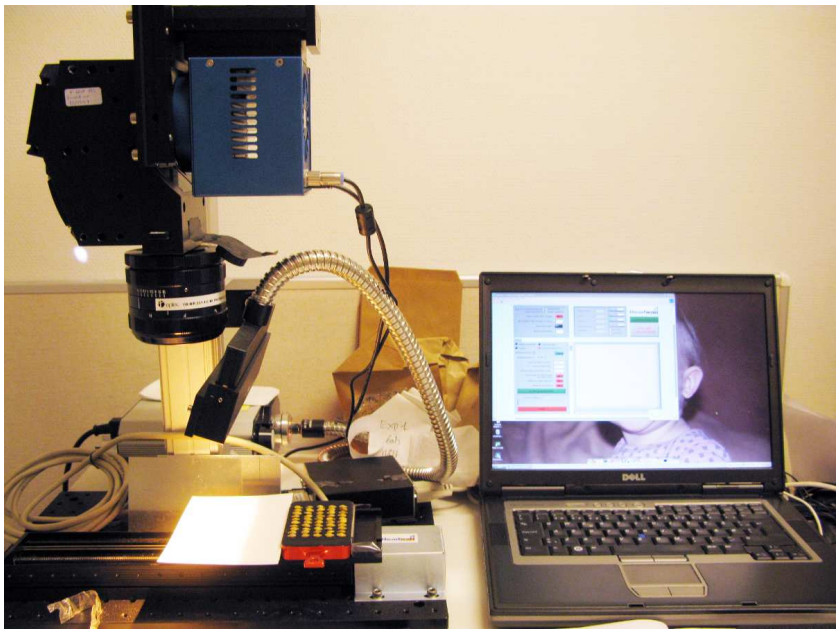


Figure 2.9: Setup of the hyperspectral line-scan NIR camera system with light source and sample conveyor belt.

## 2.6 Image and Spectral Pre-Processing

Every image acquisition procedure involves a series of pre-processing steps applied to the image- and spectral information [5, 7]. Most of the steps are mandatory, such as white reference and dark current compensation, and some are specific to the application. The most influential application specific pre-processing step is the scatter correction.

### 2.6.1 Scatter Correction

In hyperspectral image acquisition the raw captured reflectance spectra can suffer from a high degree of variation between the individual spectra as shown in Figure 2.10. This scatter effect can be described mainly due to three sources [63]

- Non-specific scatter of radiation at the surface of particles.

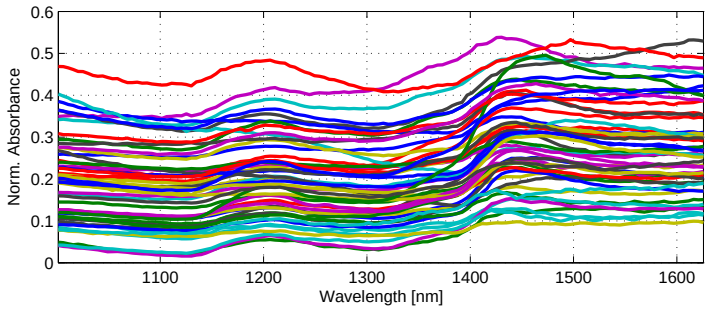



Figure 2.10: Observed raw spectra suffering from severe scattering.

- Variable spectral path length through the sample.
- Chemical composition of the sample.

During acquisition of images the intensity of the individual pixels are highly dependent on the absorption of the material and the light illumination. Objects with structured surface will not be illuminated uniformly and will lead to a scatter effect, with separate offsets for each observed spectra. This is often the major scatter effect and represents a large variance in the observed data. Figure 2.10 illustrates the scatter effects from an acquired image of barley kernels with non-uniform surface structure.

The objective is foremost to reduce any scatter effect, which may influence any subsequent analysis. A whole series of methods and approaches exist to correct for scatter effects. Rinnan *et al.* [92] makes an introduction to some of the most widely used scatter correction methods, some of which include

**Detrending** [11].
A set of basis spectra are removed from the raw data $\boldsymbol{X}$ individually either as a polynomial fit (typically 0th or 1st order) or as predefined spectra of undesired components, such as water. This approach eliminates general or common components in the spectra.

**Derivatives**.
The family of derivatives is often used to remove the mean and enhance spectral features relevant for the application at hand. They act on the individual spectra independently. Derivatives suffer from the disadvantage of undesired amplification of the background noise. However, this can be reduced by using a smoothing approach, such as the *Savitzky-Golay* approach [94].

**Multiplicative Scatter Correction (MSC)** [30, 36, 51, 76].
The MSC method attempts to reduce the scatter effects based on adjusting all observations to an 'ideal' spectrum. The mean spectrum of all observations is used as reference and all spectra are affine estimated relative to this reference. The approach makes the MSC method sensitive to the mean spectrum and thus requires re-correction for new observations.

**Standard Normal Variate (SNV)** [11, 30].
The SNV normalises each spectra individually by removing the intensity offset and scaling to unity standard deviation. The SNV is similar to MSC with different parameters. This method is very popular in many applications due to the normalisation of the spectra [5]. Unfortunately the scaling can lead to the data lying on a multidimensional sphere. This means the simplex structure in the data is lost and any unmixing becomes impossible.

The major scatter effect is often the offset of the spectra, which most scatter correction methods remove. Subtracting the mean for each spectrum is straightforward, but introduces negative spectral components and hence violates the assumption of non-negative spectra. In practice it usually never becomes a problem as adding a common offset can move the data cloud into the positive orthant. The scatter correction should ideally be chosen to remove the scatter effect present in the data. The widely accepted approach is, however, to evaluate the different scatter correction methods in conjunction with the analysis or modelling at hand. In practice the scatter correction thus often becomes application specific and acts as a feature extraction step prior to the actual analysis or modelling. Detrending and the SNV correction methods are described in details in Appendix F.

## 2.7 Summary

This introduction to hyperspectral image analysis provides a short background to the field and basic knowledge for food quality applications based on hyperspectral images. The description of three light interaction schemes give rise to a set of linear mixing models under certain conditions, which can be visually analysed and also serve as foundation for a visualisation tool for hyperspectral images.

In real applications hyperspectral image analysis can include specific processing steps, such as visualisation of the data, compression, feature extraction, modelling of references etc.

Chapter 3

# Visualisation of Hyperspectral Images of Grain

Visualisation of regular RGB images is easy using modern day hardware such as monitors. Hyperspectral images with numerous color bands cannot be viewed in the same manner and require different techniques. Initially three pre-selected bands of the hyperspectral image can be visualised at once, but only shows part of the entire image.

A linear transformation could then be chosen to find a suitable linear combination of bands to represent the predominant structures in an image. For this purpose a range of linear models exists defined as in Eq. (2.7), but with different constraints. One of the most simple transforms is the *principal component analysis (PCA)* [48], which maximises the variance in the spectral information and disregards any spatial information. Switzer *et al.* introduced the *maximum autocorrelation factor (MAF)* transform, which as the name implies, maximises the autocorrelation between neighbouring pixels based on spatial information [81, 82, 107]. A similar transform was presented by Green *et al.* as the *Minimum Noise Fraction (MNF)* transform, which maximise the signal-to-noise ratio in the image [43, 81]. The *Noise Adjusted Principal Component (NAPC)* was introduced by Lee *et al.* [60] and is mathematically equivalent to MNF. The noise covariance matrix in MNF can be based on spatial information of neigh-

bour pixels similar to MAF and these two approaches can therefore represent the same transform. Similar subspace projection approaches include the *HySime* algorithm introduced by Bioucas-Dias *et al.* [15]. Kernel versions of both the PCA and MAF/MNF transforms have been introduced by Schölkopf [97] and Nielsen [80], respectively, as non-linear variants. Larsen *et al.* [58] present an explorative study on the application of visualising the internal structure of maize kernels, where PCA, MAF/MNF and their corresponding kernel versions are compared (cf. Appendix A). These approaches all lack the interpretation of the constituent concentrations as the decompositions do not include any constraints on the fractional abundances such as Eq. (2.8). PCA and NMF are indeed suitable and widely used for visualising hyperspectral images [58], but fail to provide concentration maps of the decomposed components. Instead, unmixing is a decomposition tool for both visualising and interpreting hyperspectral images, as it requires the fractional abundances $\boldsymbol{H}$ to be non-negative and sum to one for each pixel, as described in Chapter 2 and Eq. (2.9). We will therefore focus on unmixing as the interpretation tool and present novel unmixing methods to decompose hyperspectral images.

In this section we will give a brief introduction to unmixing and discuss our novel Bayesian unmixing method based on non-negative matrix factorisation using volume priors (cf. Appendix B, C and D) in the context of visualising the constituents of wheat kernels.

## 3.1 Unmixing Introduction

The background of hyperspectral image analysis, as described in Chapter 2, gives rise to a set of mixing models depending on the scene and other factors (cf. Section 2.2). Unmixing is the process of decomposing a hyperspectral image by estimating the underlying true spectral signatures or endmembers and the associated fractional abundances.

Depending on the mixing model different unmixing methods have been developed. In Appendix B and C we give a brief introduction to some of the most popular methods for linear unmixing of hyperspectral images. Non-linear mixing is, as described, more complex and non-linear unmixing approaches have therefore not received the same attention [19, 26, 77, 79, 91, 103, 111].

The linear mixing model can under certain assumptions be considered predominant as already argued. Assuming a linear mixing of the constituents, the mixing model and constraints are expressed in Eq. (2.7) and Eq. (2.8). The unmixing of a hyperspectral image can be conducted using many different approaches.

The different unmixing methods are developed using specific applications and datasets, which are founded on a set of assumptions. In our work we have a specific set of assumptions relevant to food applications.

- **No pure pixels**
  In food applications the assumption of pure pixels in the observed data is not necessarily justified (cf. Section 2.4), as in some of the convex geometrical approaches incl. [4, 87, 89], AMEE [88], PPI [18], N-FINDR method [110], and VCA [78].

- **Number of endmembers**
  The number of endmembers $K$ is assumed known to the unmixing method and is pre-set by hand. This is a simplification, since the actual number of constituents in foods is not known. On a macroscopic scale foods are dominated by a few constituents (water, protein, starch, and fat) and $K$ therefore often lies in the range of $3-5$. Several mathematical methods exists to estimate $K$ [25, 34, 87].

- **No spectral prior information**
  The approach is unsupervised with no prior information given in the form of a weak spectral guide or a specific spectral library, which can be expensive or difficult to generate. The majority of unmixing algorithms thus infer the endmembers from the observed data. If a spectral library exists the unmixing process becomes an inversion problem, where the fractional abundances are inferred from the observations under the constraints in Eq. (2.8), eg. The Spectral Assistant (TSA) [12] or the unmixing framework presented by Iordache *et al.* incorporating a sparsity constraint on the fractional abundances to select few spectra from a given library [49, 50].

- **No sparsity**
  In food applications the data is often predominated by mixtures of constituents and not of pure pixels. Sparse unmixing may therefore not be appropriate in the analysis of food image data. Sparse approaches can be useful in conjunction with a spectral library, as mentioned.

In the light of these assumptions and the associated constraints in Eq. (2.8), the statistical approach of *Non-negative Matrix Factorization (NMF)* becomes relevant. It can estimate both the endmembers and fractional abundances without requiring pure pixels in the dataset. The non-negativity and sum-to-one constraints on the fractional abundances are rarely sufficient to capture the true endmembers in a unmixing problem. The solution is not unique as any encapsulation of all observed data points will have the same cost, ie. a non-convex cost function. Hence additional regularisation or prior knowledge is required, which

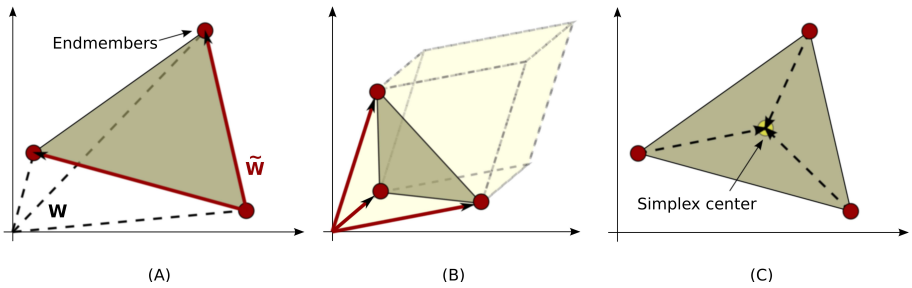



Figure 3.1: Illustration of different volume measures for a 2-simplex. (A): The true volume can be measured directly from the three endmembers $\boldsymbol{W}$ or any two difference vectors $\tilde{\boldsymbol{W}}$ (red). (B): The volume of the parallelepiped spanned by the endmembers is proportional to the volume of the encapsulated simplex. (C): The volume can be based on an Euclidean distance measure from the endmembers to the center of the simplex.

can be expressed by generalising the least-squares cost function with augmented regularisation terms,

$$C(\boldsymbol{W},\boldsymbol{H}) = \tfrac{1}{2}||\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}||^2 + \gamma J_w(\boldsymbol{W}) + \beta J_h(\boldsymbol{H}), \tag{3.1}$$

where $J_w(\boldsymbol{W})$ and $J_h(\boldsymbol{H})$ are the regularisation terms for the endmembers and fractional abundances, respectively. Different regularisations exist to improve the unmixing as described in Appendix B and C.

A recent approach is to incorporate a regularisation based on the volume of the encapsulating simplex spanned by the endmembers [14, 23, 68, 69]. A prior based on minimising the volume simplex has the advantage of encouraging a decomposition, where the estimated endmembers lie closer to the observed data. The approach does not guarantee a unique solution as the endmember identities are rotationally ambiguous. Figure 3.1 illustrates three different simplex volume measures: *Simplex volume*, *Parallelepiped*, and *Euclidean distance* (cf. Appendix B and C).

Regularisation based on the true volume of a $(K-1)$-simplex with $K$ vertices in $\boldsymbol{W}$ denoted as the column vectors $\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K\}$, where $K < M$, is expressed as

$$J_w^{\text{vol}}(\boldsymbol{W}) = \frac{1}{K!}\sqrt{\det(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}})}, \tag{3.2}$$

where the column vectors of the matrix $\tilde{\boldsymbol{W}} = [\boldsymbol{w}_1-\boldsymbol{w}_\rho\ldots\boldsymbol{w}_{\rho-1}-\boldsymbol{w}_\rho,\boldsymbol{w}_{\rho+1}-\boldsymbol{w}_\rho\ldots\boldsymbol{w}_K-\boldsymbol{w}_\rho]$ point to the vertices of the simplex from an arbitrarily selected

vertex, $\boldsymbol{w}_\rho$ (cf. Figure 3.1(A) for $K = 3$). The volume of a parallelepiped can be used as proposed by Schachtner *et al.* [95] and is expressed as

$$J_w^{\text{pp}}(\boldsymbol{W}) = \det(\boldsymbol{W}^\top \boldsymbol{W}). \tag{3.3}$$

These two volume measures are based on calculating a determinant, which is computational expensive and can potentially fail if the $K$ dimensional volume shrink to a $K-1$ subspace, where the volume becomes zero. A more robust and computationally faster measure is based on the Euclidean distance from the endmembers to the center of the simplex given by

$$J_w^{\text{dist}}(\boldsymbol{W}) = \sum_{k=1}^{K} \left\| \boldsymbol{w}_k - \frac{1}{K} \sum_{k'=1}^{K} \boldsymbol{w}_{k'} \right\|_2^2. \tag{3.4}$$

Unmixing of hyperspectral images based on the NMF model with the volume minimising regularisations in Eq. (3.2) and (3.4) has not been addressed before. Thus, we present two different frameworks, which employ the NMF model with volume minimising regularisation for unmixing.

The first algorithm is an unsupervised projected gradient NMF framework with three simplex volume measures implemented and is denoted *NMF-Vol* [10]. The algorithm employs a projected gradient approach to estimate the optimal parameters $\boldsymbol{W}$ and $\boldsymbol{H}$ as opposed to the popular multiplicative update method [59], which in certain cases can suffer from slow convergence.

Most of the unmixing algorithms are based on point estimates of the endmembers and fractional abundances, ie. *maximum likelihood (ML)* or *maximum a posteriori (MAP)*. A more general approach is to infer their posterior distributions in a Bayesian framework to provide a mean or median posterior solution with credible intervals, which are optimal under squared and linear loss respectively. In hyperspectral image analysis the credible intervals can provide additional information not possible in ML or MAP based unmixing algorithms. In a Bayesian model it is further easier to include prior information and integrate new knowledge in the form of prior distributions compared to MAP based models. Bayesian NMF models are not new and has been proposed in different contexts [31, 32, 33, 70, 71, 72, 73, 83, 96]. Our second method employs an unsupervised Bayesian framework incorporating three different simplex volume priors to estimate the joint posterior distribution of the endmembers and fractional abundances and is denoted *BayesNMF-Vol* [9].

Both methods are described in Appendix B and C with emphasis on the Bayesian framework and are available as `MATLAB` toolboxes [9, 10]. In the following sections we will present a summary of relevant details and unmixing results to highlight the performance and limitations of our unmixing methods.

## 3.2 Bayesian Inference

### 3.2.1 Introduction

In frequentist statistical methods probabilities are referred to as objective properties of the real world, eg. any modelled variance is considered a property of the observed system. Model parameters are assumed unknown constants with no useful probability statements associated to them.

A different approach is the *Bayesian Inference*, which provides a framework, where probabilities provide a quantification of uncertainty or degree of belief, eg. any modelled variance describes our ignorance of the observed system. A significant extension of the framework is how it formally incorporates prior information not provided by the observed data.

For a given set of observations $\mathcal{X}$ assume the objective is to estimate a set of model parameters $\Theta$, which best model the observed data. These parameters are considered stochastic in Bayesian inference and we wish to model their distribution conditioned on the observed data $p(\Theta|\mathcal{X})$. Bayes' theorem expresses this conditional distribution as the posterior distribution given by

$$\overbrace{p(\Theta|\mathcal{X})}^{Posterior} = \frac{\overbrace{p(\mathcal{X}|\Theta)}^{Likelihood}\overbrace{p(\Theta)}^{Prior}}{\underbrace{p(\mathcal{X})}_{Evidence}}, \tag{3.5}$$

where the prior information on the model parameters is represented in $p(\Theta)$ and the evidence ensures normalisation of the distribution. The posterior can afterwards be used to extract optimal model parameters, eg. as mean posterior or the MAP estimate. The special case of a flat prior leads to the ML estimate. A more detailed introduction to Bayesian inference can be found in [17, 21, 40].

### 3.2.2 Bayesian Unmixing

Our specific Bayesian linear unmixing model uses Bayes' rule to calculate the joint posterior of $\boldsymbol{W}$ and $\boldsymbol{H}$ expressed as

$$p(\boldsymbol{W}, \boldsymbol{H}|\boldsymbol{X}, \mathcal{H}) = \frac{p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \sigma^2)p(\boldsymbol{H})p(\boldsymbol{W}|\gamma)p(\sigma^2|\alpha, \beta)}{p(\boldsymbol{X})}, \tag{3.6}$$

where $\mathcal{H} = \{\alpha, \beta, \gamma\}$ are hyperparameters and the parameters of the model are denoted $\mathcal{P} = \{\boldsymbol{W}, \boldsymbol{H}, \sigma^2\}$. The linear mixing model is expressed by the

Gaussian distribution by assuming i.i.d. white Gaussian noise and is given by

$$p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\sigma^2)=\prod_{n=1}^{N}\prod_{m=1}^{M}\mathcal{N}\left(x_{mn}|\boldsymbol{w}_{m:}\boldsymbol{h}_{:n},\sigma^2\right), \tag{3.7}$$

where $\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x-\mu)^2}{-2\sigma^2}\right)$ is the Gaussian probability density function. The negative logarithm of the likelihood corresponds to the squared error term in Eq. (3.1). The prior of the fractional abundances follow a Dirichlet distribution with unit parameters (ie. uniformly distributed over the unit simplex, cf. Figure 2.6) to enforce the constraints in Eq. (2.8), ie.

$$p(\boldsymbol{H})\propto\prod_{n=1}^{N}\mathbb{I}\left[\sum_{k=1}^{K}h_{kn}=1\right]\prod_{k=1}^{K}\mathbb{I}\left[h_{kn}\geq 0\right], \tag{3.8}$$

where $\mathbb{I}[\cdot]$ denotes an indicator function, which has the value one when its argument is true and zero otherwise. The conjugate prior on the endmembers includes the minimum volume measure and are further assumed non-negative for proper interpretation. This is expressed as a constrained Boltzmann-Gibbs distribution

$$p(\boldsymbol{W}|\gamma)\propto e^{-\gamma J_w(\boldsymbol{W})}\prod_{m=1}^{M}\prod_{k=1}^{K}\mathbb{I}\left[w_{mk}\geq 0\right], \tag{3.9}$$

where the term $J_w(\boldsymbol{W})$ designates the chosen volume measure. Finally the noise variance $\sigma^2$ is modelled by the inverse-Gamma distribution with parameters $\alpha$ and $\beta$ given by

$$p(\sigma^2|\boldsymbol{X},\boldsymbol{W},\boldsymbol{H})=\mathcal{IG}(\sigma^2|\alpha,\beta), \tag{3.10}$$

The posterior distribution in Eq. (3.6) of the model parameters is inferred using a Markov Chain Monte Carlo (MCMC) sampling procedure since any analytical solution is impossible. Different tractable schemes exist to sample the posterior distribution, such as the Metropolis-Hastings procedure [47, 67] used in [72] or the Gibbs sampler [22, 38, 39, 41] used in [31, 32, 33, 72] (which is a special case of the Metropolis-Hastings method). The detailed derivations of the conditional distribution of the model parameters can be found in Appendix D.

The use of a sampling approach to infer the posterior is a slow process compared to point estimate methods (ML or MAP) and hence requires a higher computational load. This is the trade off to gain a posterior mean or median solution with credible intervals. Depending on shape of the posterior distribution a sampling schemes, such as the Gibbs sampler, may require numerous samples to capture the entire distribution.

The novelty of our Bayesian model is the volume regularisation, as argued. However conducting Bayesian inference will in this case lead to a intrinsic minimisation of the simplex volume. In Appendix C we derive how the posterior marginals $p(w|\boldsymbol{x})$ and $p(h_n|\boldsymbol{x})$ exhibit a polynomial decay proving the posterior mass is located in the vicinity of the observed data points, which enforces a minimal volume. Similar arguments under different derivations have been presented by Dobigeon *et al.* in [31]. Despite this intrinsic volume minimisation it can still be reasonable to enforce a minimum volume via our prior in several cases. The mixing profile of an observed data set may require additional minimisation of the simplex volume to capture the true endmembers. The influence of potential outliers can also be further suppressed with additional volume minimisation.

## 3.3 Model Initialisation

Both our unmixing algorithms, NMF-Vol and BayesNMF-Vol, relies on predefined endmember initialisation and number of endmembers $K$. In our case we initialise the endmembers $\boldsymbol{W}_{init}$ from the observations $\boldsymbol{X}$ as random samples. A more consistent approach is the maximin-distance algorithm [108], where Plaza *et al.* show improved unmixing performance [87].

Estimating the number of endmembers $K$ can be done in several ways. A common approach is to analyse the structure of the observed data from scatter plots of PCA scores. The number of PC's required to include the simplex structure can be used to indicate $K$, depending on the mixing profile and level of noise. This can be combined with prior knowledge of the application. In food applications the number of macroscopic constituents may be known and often lies in the range of $3-5$.

## 3.4 Unmixing Application

The performance of both the NMF-Vol and the BayesNMF-Vol algorithms were evaluated on a synthetic data set. In Appendix B and C a range of different number of endmembers, noise levels and dimensions were applied and show successful unmixing of the data. In comparing the two unmixing algorithms the Bayesian framework is often superior in inferring the true endmembers as they lie within the 90% credible interval (cf. Figure 3.2 right). The MAP based approach is more sensitive to a proper choice of regularisation to find endmembers, which lie within the vicinity of the true (cf. Figure 3.2 left). Simulated evidence of

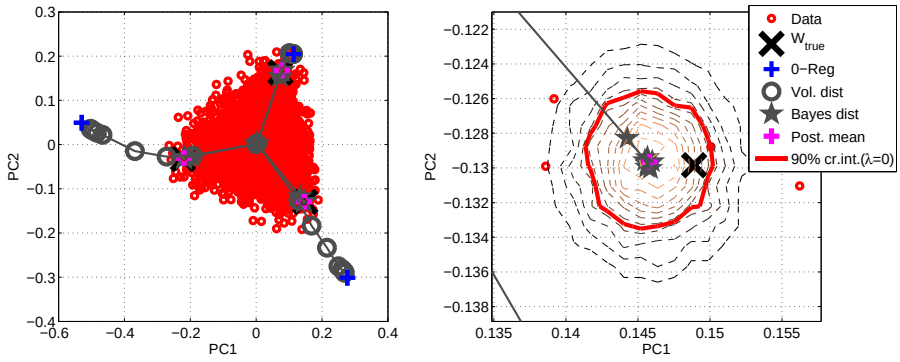



Figure 3.2: Comparison of unmixing performance of the two algorithms NMF-Vol and BayesNMF-Vol at different volume minimisation levels using the Euclidean distance measure. The left illustration shows a synthetic data set with known true endmembers, which must be inferred from the observations (zoom on the right). The MAP approach (NMF-Vol) finds endmembers in the vicinity of the true, but is sensitive to a proper choice of regularisation (left figure). The posterior means and credible intervals provided by the Bayesian framework reveal superior performance in inferring the true endmembers.

the intrinsic volume minimisation in the BayesNMF-Vol model is also shown in Figure 3.2, where increasing volume minimisation has negligible effect until a certain point.

In a practical food application the methods were also applied to visualise and interpret the internal structure of wheat kernels. A number of hyperspectral images of regular wheat kernels form the basis for the data set. In this application we assume surface penetration is dominant and that only one light path is significant per pixel (cf. Figure 2.5C). The observed spectra are hence transformed to the absorbance domain in contrast to other unmixing data set in the literature, which operate with reflectance data. Figure 3.3 illustrates an example of a single wheat kernel and a typical spectrum of a pixel. Details of the data set, pre-processing steps and unmixing are described in Appendix B and C. Figure 3.4 shows the internal constituents of a wheat kernel and the associated spectral signatures or endmembers derived as the posterior mean from unmixing using the BayesNMF-vol algorithm. The images designate the fractional abundance maps and the concentrations of the constituents can therefore be interpreted directly. The associated spectral signatures can be analysed for identification of known constituents (cf. Figure 2.3). The first component has captured a bit of the background at the periphery of the kernels. The second and fourth components represent starch and protein in two different environments, which for

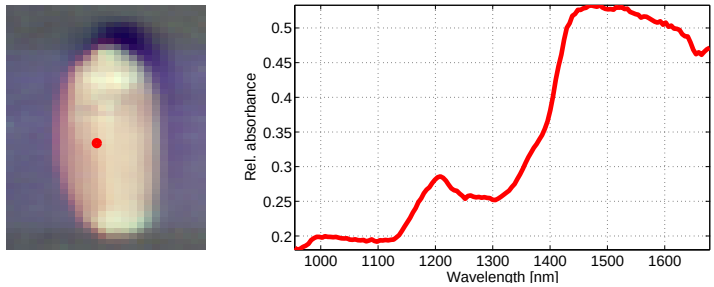



Figure 3.3: Pseudo RGB image of a single wheat kernel and a typical spectra of a pixel.

wheat kernels lie in a matrix and are difficult to separate due to their correlation. The germ of the kernel contains fat and is partially captured by the third component.

The hyperspectral image technology offers data analysis on a single kernel level. For the wheat kernel data set the protein levels are one of the indicators of the food quality. These ground truth concentrations were not available and thus this data set lacks a set of references and subsequent quality assessment. In a practical application with the objective of building a model for measuring the protein levels in wheat kernels, a range of kernels with different protein levels and corresponding references should be available.

## 3.5 Summary

Our two unsupervised unmixing methods proved successful in decomposing both synthetic data and wheat kernel data into interpretable components. Extending the regular NMF model with a volume minimising regularisation improved the unmixing performance in the NMF-Vol method and reduced the sensitivity towards noise and mixing profile as opposed to the convex geometry unmixing types, eg. [18, 78, 88, 110]. In the Bayesian framework the volume minimisation is intrinsic in the likelihood term and thus the volume minimising prior is not strictly necessary. The Euclidean distance based volume measure was further slightly superior in computational load and in its reduced sensitivity to linear dependencies among the endmembers (cf. Section 3.1 and Appendix C). The validation of the unmixing performance on the wheat kernel data suffered from the lack of true protein concentrations for the single kernels, and forced a subjective evaluation.

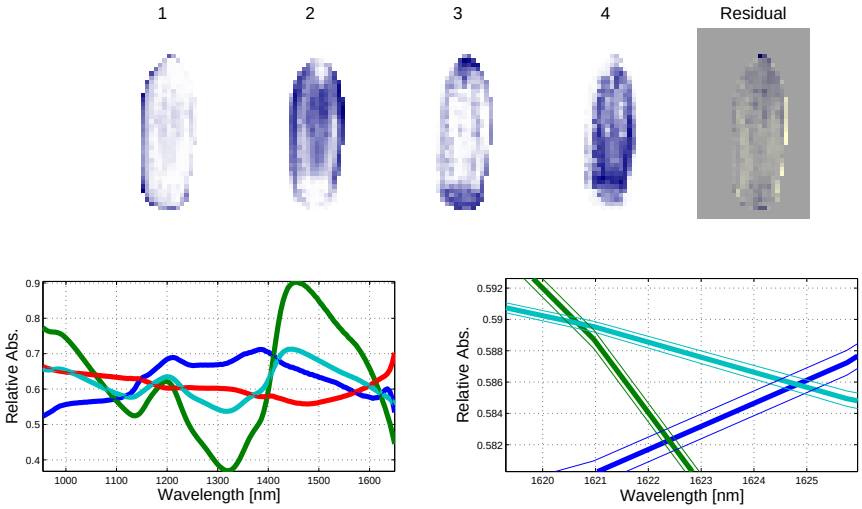



Figure 3.4: Image components of an unmixed wheat kernel as the posterior mean from the BayesNMF-Vol algorithm. The associated spectral signatures or endmembers are shown below with a zoom showing the 95%-quantile credible interval.

Our algorithms comprise an open implementation and can easily be extended in many ways. Currently the strength of volume minimisation parameter $\gamma$ is pre-set by hand. An optimal $\gamma$ can easily be estimated using a cross validation approach or in the Bayesian model treated as a hyperparameter in a hierarchical framework. In estimating the endmembers any highly mixed pixels in the middle of the simplex have negligible contribution. For speed improvements these pixels can be omitted and decomposed when the endmembers are found, eg. by a simple algorithm developed by Mørup *et al.* [75], which hollows the simplex. The algorithms can further be applied to different mixing profiles of synthetic data sets to evaluate their robustness.

In this food application we assumed subsurface penetration is not negligible and therefore linearised the reflectance data into the absorbance domain. In our case we could not validate this assumption due to the lack of ground truth references. Hence, a proper approach is to validate such an assumption empirically by a data set with known true references (endmembers or fractional abundances). This quantitative evaluation applies to the different mixing models described in Section 2.2 incl. non-linear mixing contribution.

The unmixing method does not apply to single spectrum NIR systems and thus illustrates the immediate advantage of extending the system to include spatial

information. Multispectral images can easily be subjected to unmixing as well, but may lack the spectral interpretation of the endmembers or spectral signatures compared to the abundance of spectral bands provided in hyperspectral images. For food quality assessments we have shown how unmixing hyperspectral data can be used to visualise and interpret the internal structure of the food and also serve as an analysing tool for the quality assessment.

Chapter 4

# Prediction of Pre-Germination in Barley

Hyperspectral image analysis can be used to directly model any labels associated to the grain kernels instead of visualising their internal structure. Such an approach is supervised by definition and is conducted with the objective of extracting features, which are optimal to the specific application for best performance and not with respect to interpretation as in the unmixing method described in Chapter 3.

Using hyperspectral image analysis to predict labels in foods is not a new discipline. In Appendix E we give a brief introduction to related research. In this section we discuss a practical food quality assessment application to predict the degree of pre-germination in barley kernels (cf. Appendix E). A modelling framework customised to this particular application is presented.

## 4.1 Introduction

Barley kernels (Hordeum vulgare) is a member of the grass family and is an important food resource in many areas, eg. animal fodder or base malt for beer

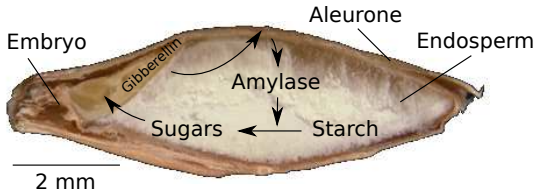



Figure 4.1: Illustration of the internal structure of a barley kernel and the biochemical mechanisms during germination (*Source: KAMPFFMEYER Food Innovation GmbH*). The germination process begins with the absorption of water leading to the release of the enzyme $\alpha$-amylase from the aleurone layer into the endosperm. The enzyme breaks the starch molecules down into the disaccharide maltose and finally glucose, which is transported to the embryo for radicle and epicotyl growth.

and certain distilled beverages. The quality of barley can be described from many different parameters. One of the more significant parameters is the degree of pre-germination, which has a particular high negative impact for the malting industry. The quality assessment of barley in terms of pre-germination is in our case to identify the degree of pre-germination in single kernels.

Germination of barley is the process of the kernel starting to sprout and becoming a plant. It involves a whole series of biochemical steps, where the enzyme $\alpha$-amylase breaks the starch down into maltose (cf. Figure 4.1). The progression illustrated in Figure 4.2 shows the chemical changes inside a kernel during germination. It starts with a short period with little activity, where the $\alpha$-amylase release is commencing and rises as the amylase activity is maximum and eventually settles as all the starch is broken down. A general introduction to barley kernels and the germination process can be found in Appendix E.

Current standard procedures to identify pre-germination of barley involves destructive and labour intensive manual procedures [2, 3]. The germination process is seen as a large scale physical change inside the kernel and can be captured by a hyperspectral NIR camera system. Thus, hyperspectral image analysis has the potential to provide a fast and non-destructive analysis as an alternative approach to identify pre-germination. It further offers the opportunity to exploit spatial features as opposed to single spectrum NIR systems.

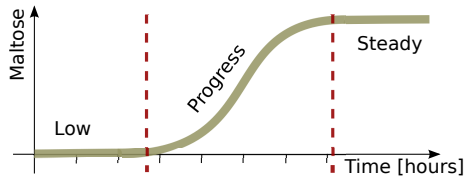



Figure 4.2: Conceptual illustration of the progression of germinating barley showing the starch break down. The two thresholds indicate the start and stop of the major chemical changes inside the kernel during germination.

## 4.2 Pre-germination Application

A dedicated pre-germination experiment was set up to form a hyperspectral image data set used to investigate the potential of using hyperspectral image analysis to identify pre-germination in barley. A mathematical modelling framework was build from the data set with the objective to classify single kernels into different degrees of pre-germination. Barley kernels with eight different pre-germination times: $[0, 12, 18, 24, 30, 36, 48, 60]$ hours were generated. Hyperspectral images were acquired of a subset of the kernels to comprise the data set and other kernels were re-germinated to evaluate the negative impact of pre-germination using the BRF-method [1]. The re-germination gave rise to dividing the impact of pre-germination into three groups identified as

1. **Normal re-germination $(0 - 18\text{h})$**. *A short pre-germination has negligible effect as all kernels re-germinate as normal kernels.*
2. **Delayed re-germination $(24 - 36\text{h})$**. *The kernels will re-germinate within five days at a slower rate.*
3. **Limited re-germination $(36-\text{h})$**. *The kernels are exhausted from the pre-germination and not all will re-germinate within five days.*

This suggests that the division of pre-germination times into these three classes is optimal and not into a finer grade of our eight classes. This means that the pre-germination times 0, 12, and 18h all represent a single biological state of the kernels and not three separate classes. Similar arguments apply to $24 - 36$h and $48 - 60$h of pre-germination. Details of both the pre-germination and re-germination experiment are described in Appendix E along with the image acquisition procedure and subsequent pre-processing of the data. The processed hyperspectral image data set comprises of 755 kernels divided unequally among the eight classes, where the smallest class has 41 kernels, the largest has 135 kernels, and all kernels have an equal water content of approx. 10%.

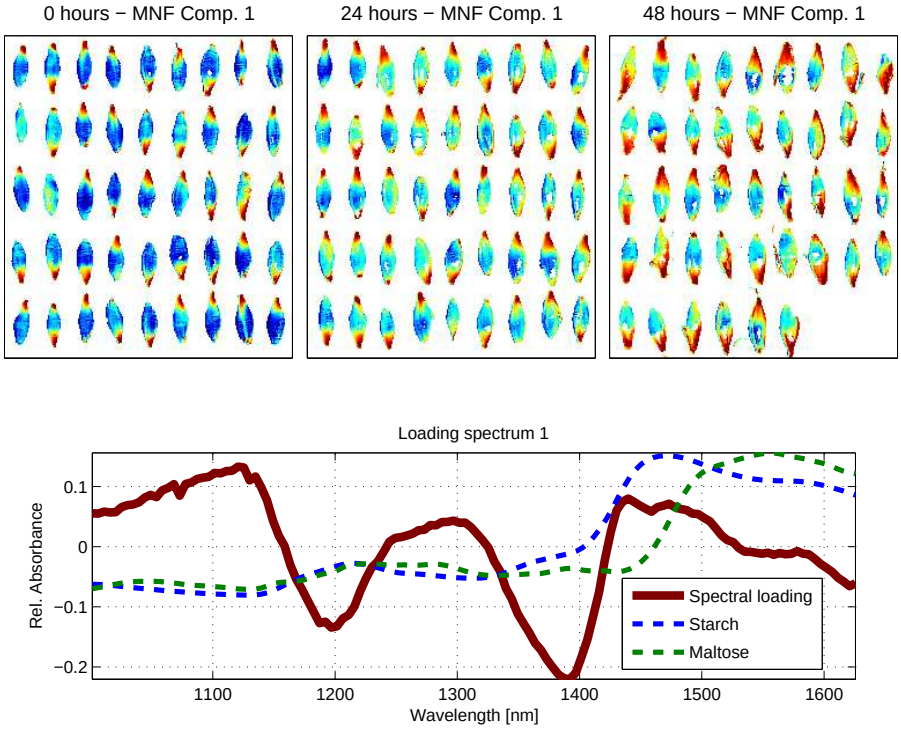



Figure 4.3: Score images from an MNF decomposition (cf. Appendix F) of the germination times $[0, 24, 48]$ (left to right), which reveal a progression in the germination process. The associated spectrum represents the correlations between the MNF feature vector 1 and the original variables and exhibits spectral activity in the area, where the spectra of starch and maltose differ.

Grain kernels have only a few basic macroscopic compounds, which relate to the germination process, namely starch and maltose. These two constituents have distinct spectral signatures in their pure form as shown in Figure 2.3. The germination process involves the break down of starch and a transport of the resulting maltose to the embryo for growth. This physical process can be seen as a growing area at the embryo of the kernels as the germination progresses. These spatial and spectral changes can be captured by the hyperspectral camera system as illustrated in Figure 4.3.

One of the crucial steps in classifying each kernel is to extract features, which captures the germination progression. Since the germination process can be observed as a physical process, the features are based on spatial information. These features are used to compress the image data into fewer variables, which can

be used for classification of single kernels. A flowchart of the feature extraction and classification framework can be found in Appendix E.

A range of different processing steps were evaluated to find the optimal framework in terms of lowest classification error. This gave rise to the evaluation of different known methods of pre-processing, feature extraction, and classification. The framework with the lowest classification error is described in Appendix E and is available as a `MATLAB` toolbox [6]. Other evaluated approaches are described in the supplementary material in Appendix F. In the following sections we give an overview of some of the feature extraction methods and classification models considered in the framework.

### 4.2.1 Feature Extraction and Classification

To extract the spatial features for enhancing the germination progression our framework rely on simple transformations, namely PCA, MAF/MNF and kernel PCA (cf. Chapter 3). The resulting components represented as monochrome score images from these three methods are presented in Appendix E and F.

In general the MAF/MNF decomposition is well suited for image data as it includes spatial information opposed to PCA and kPCA, which only consider the spectral information. The noise covariance matrix in MAF/MNF is estimated from neighbour pixel information and will hence include high frequency spatial information, which will tend to extract smoother image components compared to PCA and kPCA in this application. The spatial features are quantified in a 2nd level of feature extraction, as illustrated in Figure 4.4, which leaves each kernel $n$ represented by a 16 dimensional vector denoted $\boldsymbol{F}_n$ (cf. Appendix E).

The classification network predicts the degree of pre-germination by classifying each kernel into one of the eight classes $[0, 12, 18, 24, 30, 36, 48, 60]$ hours and three classes $[0-18, 24-36, 48-60]$ hours based on the single kernel feature $\boldsymbol{F}_n$. In our analysis we have considered different approaches for the classification, namely ordinal-, linear-, or non-linear classification, which are described below.

Regular classifiers consider each class separately and find discriminants, which best separates the classes. The germination times in our data set has an ordering, which can be exploited by ordinal classifiers. A widely used approach to incorporate ordinality is to predict the labels first using a regression model and use the predictions for the actual classification. Regression models incorporate a cost function, where predictions far from the ground truth are penalized hard, eg. least-squares loss function. This encourages misclassifications to occur to neighbour classes only. In Figure 4.5 the predictions from a non-linear neural

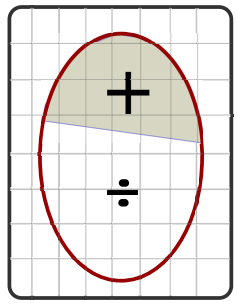


Figure 4.4: The quantification consist of two scalar values for each score image calculated as the sum of the positive and negative intensities respectively relative to the total number of pixels.

network regression model applied to our data set is shown (cf. Appendix F for different ordinal classifiers applied to our data set). The basic structure of the germination times is captured despite the large prediction variance. The similar predictions of the last two classes 48h and 60h suggest the germination process decreases after 48h of germination (cf. Figure 4.2). Conducting single kernel classification using the predictions gives an error of 33% ($CI_{95\%}$ : $30\% - 36\%$).

The regular linear multinominal classifier is a more simple method, which does not consider the ordinality in the data set. In Appendix E the multinominal classifier is applied to our data set and achieves a single kernel classification error of 32% ($CI_{95\%}$ : $29\% - 35\%$), which is comparable to the ordinal classifier. Preliminary results from a non-linear ANN classifier indicate no improved performance compared to the linear classification models.

Current standards provide quality assessment based on several kernels, eg. 100 kernels in the BRF method [1]. Similar quality measures at bulk level is easily produced from the single kernel predictions by averaging the class probabilities associated to each kernel. The ordinal and multinominal classifier reduce the error to 12% ($CI_{95\%}$ : $3\% - 28\%$) and 3% ($CI_{95\%}$ : $0\% - 15\%$), respectively, for the three classes by averaging 20 kernels as shown Figure 4.5. Both the associated error distributions reveal misclassifications in the $24 - 36$ hour pre-germination period, where the germination activity is highest. In this phase the kernels can have individual rates of germination, which means the associated labels of pre-germination times may be inaccurate leading to misclassifications.

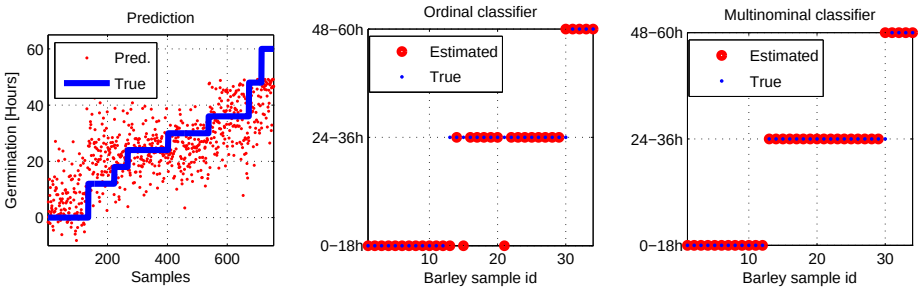



Figure 4.5: *Left:* The predicted and true germination time for each of the 755 kernels. *Middle and right:* Classification error distribution of the three classes for both the ordinal- and multinominal classifier for an average of 20 kernels.

## 4.3 Discussion

The re-germination experiment suggests the degree of pre-germination can be divided into three underlying classes. This indicates the eight classes used in our experiment represents too many degrees and will hence lead to increased classification errors. The three classes further implies a prediction of pre-germination based on a continuous variable is not feasible as such an approach provides predictions below the accuracy of the labels.

The data set consists of 755 kernel, where the smallest class is represented by 41 kernels. A future experiment must be based on more kernels to achieve classifications error with improved confidence intervals. Such pre-germinated barley kernels should preferably contain different water levels across the single kernels to limit any sensitivity towards water concentrations.

The three classes are currently modelled by aggregating the class probabilities estimated for all eight classes. A more direct approach is to retrain the model with the three classes to obtain a more dedicated model with potentially lower classification error.

## 4.4 Summary

Using hyperspectral image analysis in a classification framework proved a success in prediction of the degree of pre-germination for single barley kernels. The mathematical models used are all known and well-established and does not represent any innovation. The novelty is represented by the framework in the

application of pre-germinated barley. The classification framework is also applicable to other quality parameters in single grain kernels, such as estimating the degree of surface contaminations.

The NIR hyperspectral image system can penetrate kernels and capture spatial changes occurring during the germination. This progression is exploited by our classification framework as spatial features. A single spectrum NIR system can therefore not perform similar quality assessment, but must rely on average spectra instead.

A re-germination experiment revealed three underlying classes of pre-germination. These three classes were modelled in our classification framework based on a multinomial classifier and achieved a 32% ($CI_{95\%} : 29\% - 35\%$) error rate for single kernels and 3% ($CI_{95\%} : 0\% - 15\%$) for averages of 20 kernels.

Chapter 5

# Conclusion

Assessing the quality of foods is a vital discipline in all aspects from the farmer over the manufacturer to the consumer. The practical use of single spectrum NIR systems to conduct a fast and reliable evaluation of food quality has already been established. The augmentation of NIR analysis with image data in assessing the quality of food has given rise to new opportunities in food analysis. New quality parameters can be evaluated based on spatial features, not addressable by single spectrum NIR systems e.g. low concentration surface contaminations of grain. This has called for new modelling strategies to exploit the added spatial information into new quality assessment applications.

We have demonstrated how this additional image information can be exploited for assessing the quality of foods. Knowledge about the internal structure of food can be particularly helpful in assessing food quality. Visualising hyperspectral images can be done in many ways. Most of the decompositions disregard the sum-to-one constraint of the fractional abundances for proper interpretation, e.g. PCA and MNF. Unmixing is a decomposition approach used in hyperspectral image analysis to address this issue and hence represents part of the research contribution.

In our introduction of spectroscopy we discussed several light/matter interaction models and a set of scene dependent assumptions and simplifications, which

lead to linear mixing models for the different scenes. We presented a novel linear unmixing approach based on different minimum volume constraints in both a MAP and a Bayesian framework applied to absorbance data of hyperspectral images of food. The Bayesian model achieved better unmixing performance as the posterior mean estimates and provided credible intervals as part of the decomposition, which can act as a confidence measure. In designing multispectral NIR systems in future food quality applications this can be used to identify the most significant wavelengths. The Bayesian model further had minor intrinsic volume minimisation compared to the MAP model and can be extended manually when required by the dataset at hand. In the practical application of visualising single wheat kernels both algorithms proved successful in finding interpretable spatial and spectral structures in the data. The underlying food constituents were identified with associated concentration maps for each component in every single wheat kernel. The validation of the unmixing performance suffered from the lack of true protein concentrations for the single kernels, which forced a subjective evaluation.

In the case of single grain kernels the reference methods to measure quality parameters are mostly conducted on a bulk level. Hyperspectral image analysis offers evaluation of smaller food sample down to individual grains compared to single spectrum NIR system. This emphasises the challenge of obtaining corresponding references on a similar level.

The spectroscopic models comprise a major challenge in hyperspectral image analysis. This is particularly evident in laboratory size imaging equipment used in eg. food quality applications, where complex light/matter interactions phenomena become more dominant. The linear mixing model assumed by most unmixing algorithms is not entirely correct, but is accepted due to its simplicity and reasonable results. Improved unmixing can be achieved by considering more complex spectroscopic models and constitute a challenge in future research in unmixing.

In the second approach the practical use of hyperspectral image analysis as a food quality assessor was demonstrated. In the application we focused on acquiring reference labels for single kernel and building a customised classification framework for estimating degree of pre-germination in single barley kernels. The experiment to induce pre-germination into the barley kernels revealed three classes of pre-germination, which were not seen before. The framework exploits both the spectral and spatial data to extract area specific spectral features related to the pre-germination quality parameter. This combination leads to a strong feature, which is not possible to extract by single spectrum NIR systems, where a spectral average is used. The mathematical framework achieved a prediction of the degree of pre-germination in single barley kernels 3% ($\mathrm{CI}_{95\%} : 0\% - 15\%$) for averaging 20 kernels.

Both approaches utilise the additional information provided in the hyperspectral data and demonstrate how new mathematical models can be exploited in food quality applications, not possible by single spectrum NIR systems. Our modelling frameworks can form the basis for further development and can be exploited in new food quality applications. Both approaches are further available as `MATLAB` toolboxes [6, 9, 10].

In summary we have illustrated how the added image information in hyperspectral image analysis offers additional possibilities in new food quality assessment application. Our research demonstrates how these opportunities can be exploited in mathematical frameworks in practical food quality assessments. Hyperspectral NIR systems in general have a great potential to replace labour intensive and destructive reference methods in assessing food quality, but brings challenges such as complex spectroscopic models and acquiring reference labels, which must be addressed.

## 5.1 Perspective

Our unmixing model and classification framework are two examples illustrating how mathematical models can be built to exploit hyperspectral imaging in food quality assessments. They show a path for further development, but only represent a small potential in hyperspectral imaging. The future capabilities of hyperspectral imaging can be further supported with new advanced mathematical models to explore new opportunities in this field.

One approach is to focus on sophisticated non-linear unmixing algorithms to meet the challenges of more complex spectroscopic models and include scene dependencies, eg. [79], cf. Section 2.2.

The unsupervised unmixing method is concerned with finding interpretable features and does not model any reference labels. In contrast the classification framework is focused on modelling the pre-germination quality parameter directly with an optimal set of features, which may not necessarily be interpretable. In essence it often becomes a trade-off between interpretation of the data and optimal modelling of specific labels.

A natural extension to both models is to combine them into a framework, where the single kernel references are modelled from interpretable features, or in other words to model the aforementioned trade-off. A class of models, which addresses this trade-off are the *Latent Variable Multivariate Regression models (LVMR)* [20], which include the Partial Least-Squares models (PLS) [37, 65]. A general

LVMR model can easily be extended to include the non-negativity and sum to one constraint of the fractional abundances. Such customised LVMR model can perform linear unmixing to extract a set of interpretable features used to predict the labels. The concept is to let the predictions of the labels affect the unmixing to encourage spectral endmembers, which capture the spectral information correlated to the labels. In the pre-germination application, this means to encourage endmembers describing the pre-germination progression. Such an approach further involves multiple instance learning, since references are associated to whole grain kernels and not the individual pixels.

In terms of food quality applications hyperspectral imaging can easily be applied to other types of foods, such as grapes, meat or dairy products.

Appendix A

# Kernel based subspace projection of near infrared hyperspectral images of maize kernels

Authors: Rasmus Larsen, Morten Arngren, Per Waaben Hansen and Allan Aasbjerg Nielsen.

Kernel based subspace projection of near infrared hyperspectral images of maize kernels. *Lecture Notes in Computer Science, vol. 5575/2009, p. 560-569, 2009.* Published.

# Kernel Based Subspace Projection of Near Infrared Hyperspectral Images of Maize Kernels

Rasmus Larsen[1], Morten Arngren[1,2], Per Waaben Hansen[2], and Allan Aasbjerg Nielsen[3]

[1] DTU Informatics, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
{rl,ma}@imm.dtu.dk
[2] FOSS Analytical AS, Slangerupgade 69, DK-3400 Hillerød, Denmark
pwh@foss.dk
[3] DTU Space, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
aa@space.dtu.dk

**Abstract.** In this paper we present an exploratory analysis of hyperspectral 900-1700 nm images of maize kernels. The imaging device is a line scanning hyper spectral camera using a broadband NIR illumination. In order to explore the hyperspectral data we compare a series of subspace projection methods including principal component analysis and maximum autocorrelation factor analysis. The latter utilizes the fact that interesting phenomena in images exhibit spatial autocorrelation. However, linear projections often fail to grasp the underlying variability on the data. Therefore we propose to use so-called kernel version of the two afore-mentioned methods. The kernel methods implicitly transform the data to a higher dimensional space using non-linear transformations while retaining the computational complexity. Analysis on our data example illustrates that the proposed kernel maximum autocorrelation factor transform outperform the linear methods as well as kernel principal components in producing interesting projections of the data.

## 1 Introduction

Based on work by Pearson [1] in 1901, Hotelling [2] in 1933 introduced principal component analysis (PCA). PCA is often used for linear orthogonalization or compression by dimensionality reduction of correlated multivariate data, see Jolliffe [3] for a comprehensive description of PCA and related techniques.

An interesting dilemma in reduction of dimensionality of data is the desire to obtain simplicity for better understanding, visualization and interpretation of the data on the one hand, and the desire to retain sufficient detail for adequate representation on the other hand.

Schölkopf et al. [4] introduce kernel PCA. Shawe-Taylor and Cristianini [5] is an excellent reference for kernel methods in general. Bishop [6] and Press et al. [7] describe kernel methods among many other subjects.

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 560–569, 2009.

The kernel version of PCA handles nonlinearities by implicitly transforming data into high (even infinite) dimensional feature space via the kernel function and then performing a linear analysis in that space.

The maximum autocorrelation factor (MAF) transform proposed by Switzer [11] defines maximum spatial autocorrelation as the optimality criterion for extracting linear combinations of multispectral images. Contrary to this PCA seeks linear combinations that exhibit maximum variance. Because the interesting phenomena in image data often exhibit some sort of spatial coherence spatial autocorrelation is often a better optimality criterion than variance. A kernel version of the MAF transform has been proposed by Nielsen [10].

In this paper we shall apply kernel MAF as well as kernel PCA and ordinary PCA and MAF to find interesting projections of hyperspectral images of maize kernels.

## 2 Data Acquisition

A hyperspectral line-scan NIR camera from Headwall Photonics sensitive from 900-1700nm was used to capture the hyperspectral image data. A dedicated NIR light source illuminates the sample uniformly along the scan line and an advanced optic system developed by Headwall Photonics disperses the NIR light onto the camera sensor for acquisition. A sledge from MICOS GmbH moves the sample past the view slot of the camera allowing it to acquire a hyperspectral image. In order to separate the different wavelengths an optical system based on the Offner principle is used. It consists of a set of mirrors and gratings to guide and spread the incoming light into a range of wavelengths, which are projected onto the InGaAs sensor.

The sensor has a resolution of 320 spatial pixels and 256 spectral pixels, i.e. a physical resolution of $320 \times 256$ pixels. Due to the Offner dispersion principle (the convex grating) not all the light is in focus over the entire dispersed range. This means that if the light were dispersed over the whole 256 pixel wide sensor the wavelengths at the periphery would be out of focus. In order to avoid this the light is only projected onto 165 pixels instead and the top 91 pixels are disregarded. This choice is a trade-off between spatial sampling resolution and focus quality of the image.

The camera acquires 320 pixels and 165 bands for each frame. The pixels are represented in 14 bit resolution with 10 effective bits In Fig. 1 average spectra for a white reference and dark background current images are shown. Note the limited response in the 900-950 nm range.

Before the image cube is subjected to the actual processing a few preprocessing step are conducted. Initially the image is corrected for the reference light and dark background current. A reference and dark current image are acquired and the mean frame is applied for the correction. In our case the hyperspectral data are kept as reflectance spectra throughout the analysis.

562 R. Larsen et al.

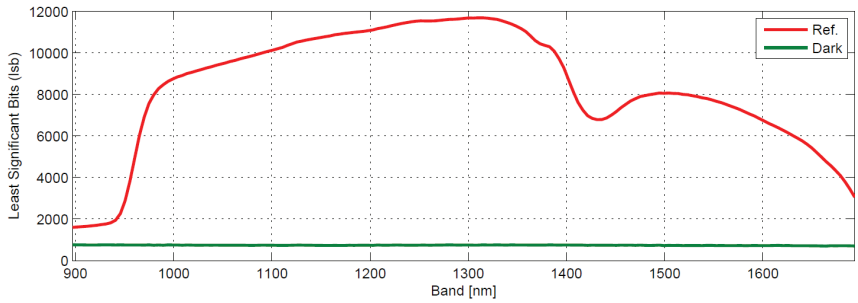
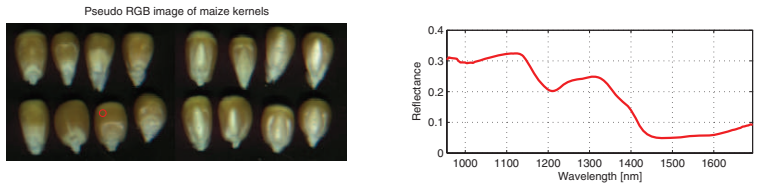



**Fig. 1.** Average spectra for white reference and dark background current images

### 2.1 Grain Samples Dataset

For the quantitative evaluation of the kernel MAF method a hyperspectral image of eight maize kernels is used as the dataset. The hyperspectral image of the maize samples are comprised of the front and back-side of the kernels on a black background (NCS-9000) appended as two separate cropped images as depicted in Fig. 2(a). In Fig. 2(b) an example spectrum is shown. The kernels are not




(a) (b)

**Fig. 2.** (a) Front (left) and back (right) images of eight maize kernels on a dark background. The color image is constructed as an RGB combination of NIR bands 150, 75, and 1; (b) reflectance spectrum of the pixel marked with red circle in (a).

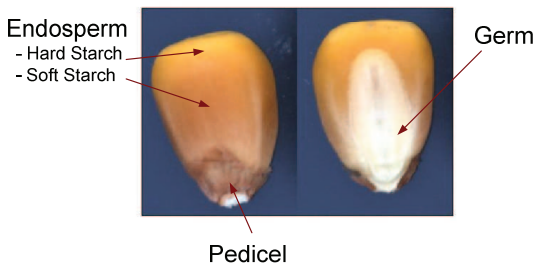



**Fig. 3.** Maize kernel constituents front- and backside (pseudo RGB)

fresh from harvest and hence have a very low water content and are in addition free from any infections. Many cereals in general share the same compounds and basic structure. In our case of maize a single kernel can be divided into many different constituents on the macroscopic level as illustrated in Fig. 3.

In general, the structural components of cereals can be divided into three classes denoted *Endosperm*, *Germ* and *Pedicel*. These components have different functions and compounds leading to different spectral profiles as described below.

*Endosperm.* The endosperm is the main storage for starch (∼66%), protein (∼11%) and water (∼14%) in cereals. Starch being the main constituent is a carbohydrate and consists of two different glucans named *Amylose* and *Amylopectin*. The main part of the protein in the endosperm consists of *zein* and *glutenin*. The starch in maize grains can be further divided into a soft and a hard section depending on the binding with the protein matrix. These two types of starch are typically mutually exclusive, but in maize grain they both appear as a special case as also illustrated in figure 3.

*Germ.* The germ of a cereal is the reproductive part that germinates to grow into a plant. It is the embryo of the seed, where the scutellum serves to absorb nutrients from the endosperm during germination. It is a section holding proteins, sugars, lipids, vitamins and minerals [13].

*Pedicel.* The pedicel is the flower stalk and has negligible interest in terms of production use. For a more detailed description of the general structure of cereals [12].

## 3 Principal Component Analysis

Let us consider an image with $n$ observations or pixels and $p$ spectral bands organized as a matrix $X$ with $n$ rows and $p$ columns; each column contains measurements over all pixels from one spectral band and each row consists of a vector of measurements $x_i^T$ from $p$ spectral bands for a particular observation $X = [x_1^T x_2^T \dots x_n^T]^T$. Without loss of generality we assume that the spectral bands in the columns of $X$ have mean value zero.

### 3.1 Primal Formulation

In ordinary (primal also known as R-mode) PCA we analyze the sample variance-covariance matrix $S = X^T X/(n-1) = 1/(n-1)\sum_{i=1}^{n} x_i x_i^T$ which is $p$ by $p$. If $X^T X$ is full rank $r = \min(n, p)$ this will lead to $r$ non-zero eigenvalues $\lambda_i$ and $r$ orthogonal or mutually conjugate unit length eigenvectors $u_i$ ($u_i^T u_i = 1$) from the eigenvalue problem

$$\frac{1}{n-1} X^T X u_i = \lambda_i u_i. \tag{1}$$

We see that the sign of $u_i$ is arbitrary. To find the principal component scores for an observation $x$ we project $x$ onto the eigenvectors, $x^T u_i$. The variance of these

scores is $u_i^T S u_i = \lambda_i u_i^T u_i = \lambda_i$ which is maximized by solving the eigenvalue problem.

### 3.2 Dual Formulation

In the dual formulation (also known as Q-mode analysis) we analyze $XX^T/(n-1)$ which is $n$ by $n$ and which in image applications can be very large. Multiply both sides of Equation 1 from the left with $X$

$$\frac{1}{n-1} XX^T(Xu_i) = \lambda_i(Xu_i) \qquad \text{or} \qquad \frac{1}{n-1} XX^T v_i = \lambda_i v_i \tag{2}$$

with $v_i$ proportional to $Xu_i$, $v_i \propto Xu_i$, which is normally not normed to unit length if $u_i$ is. Now multiply both sides of Equation 2 from the left with $X^T$

$$\frac{1}{n-1} X^T X(X^T v_i) = \lambda_i(X^T v_i) \tag{3}$$

to show that $u_i \propto X^T v_i$ is an eigenvector of $S$ with eigenvalue $\lambda_i$. We scale these eigenvectors to unit length assuming that $v_i$ are unit vectors $u_i = X^T v_i / \sqrt{(n-1)\lambda_i}$.

We see that if $X^T X$ is full rank $r = \min(n, p)$, $X^T X/(n-1)$ and $XX^T/(n-1)$ have the same $r$ non-zero eigenvalues $\lambda_i$ and that their eigenvectors are related by $u_i = X^T v_i / \sqrt{(n-1)\lambda_i}$ and $v_i = Xu_i / \sqrt{(n-1)\lambda_i}$. This result is closely related to the Eckart-Young [8,9] theorem.

An obvious advantage of the dual formulation is the case where $n < p$. Another advantage even for $n \gg p$ is due to the fact that the elements of the matrix $G = XX^T$, which is known as the Gram[1] matrix, consist of inner products of the multivariate observations in the rows of $X$, $x_i^T x_j$.

### 3.3 Kernel Formulation

We now replace $x$ by $\phi(x)$ which maps $x$ nonlinearly into a typically higher dimensional feature space. The mapping by $\phi$ takes $X$ into $\Phi$ which is an $n$ by $q$ ($q \geq p$) matrix, i.e. $\Phi = [\phi(x_1)^T \phi(x_2)^T \ldots \phi(x_n)^T]^T$ we assume that the mappings in the columns of $\Phi$ have zero mean. In this higher dimensional feature space $C = \Phi^T \Phi/(n-1) = 1/(n-1) \sum_{i=1}^{n} \phi(x_i)\phi(x_i)^T$ is the variance-covariance matrix and for PCA we get the primal formulation $1/(n-1)\Phi^T \Phi u_i = \lambda_i u_i$ where we have re-used the symbols $\lambda_i$ and $u_i$ from above. For the corresponding dual formulation we get re-using the symbol $v_i$ from above

$$\frac{1}{n-1} \Phi \Phi^T v_i = \lambda_i v_i. \tag{4}$$

As above the non-zero eigenvalues for the primal and the dual formulations are the same and the eigenvectors are related by $u_i = 1/(\sqrt{(n-1)\lambda_i})\ \Phi^T v_i$, and $v_i = \Phi\ u_i / \sqrt{(n-1)\lambda_i}$. Here $\Phi\Phi^T$ plays the same role as the Gram matrix above and has the same size, namely $n$ by $n$ (so introducing the nonlinear mappings in $\phi$ does not make the eigenvalue problem in Equation 4 bigger).

[1] Named after Danish mathematician Jørgen Pedersen Gram (1850-1916).

**Kernel Substitution.** Applying kernel substitution also known as the kernel trick we replace the inner products $\phi(x_i)^T\phi(x_j)$ in $\Phi\Phi^T$ with a kernel function $\kappa(x_i, x_j) = \kappa_{ij}$ which could have come from some unspecified mapping $\phi$. In this way we avoid the explicit mapping $\phi$ of the original variables. We obtain

$$Kv_i = (n-1)\lambda_i v_i \tag{5}$$

where $K = \Phi\Phi^T$ is an $n$ by $n$ matrix with elements $\kappa(x_i, x_j)$. To be a valid kernel $K$ must be symmetric and positive semi-definite, i.e., its eigenvalues are non-negative. Normally the eigenvalue problem is formulated without the factor $n-1$

$$Kv_i = \lambda_i v_i. \tag{6}$$

This gives the same eigenvectors $v_i$ and eigenvalues $n-1$ times greater. In this case $u_i = \Phi^T v_i/\sqrt{\lambda_i}$ and $v_i = \Phi u_i/\sqrt{\lambda_i}$.

**Basic Properties.** Several basic properties including the norm in feature space, the distance between observations in feature space, the norm of the mean in feature space, centering to zero mean in feature space, and standardization to unit variance in feature space, may all be expressed in terms of the kernel function without using the mapping by $\phi$ explicitly [5,6,10].

**Projections onto Eigenvectors.** To find the kernel principal component scores from the eigenvalue problem in Equation 6 we project a mapped $x$ onto the primal eigenvector $u_i$

$$\begin{aligned}\phi(x)^T u_i &= \phi(x)^T\Phi^T v_i/\sqrt{\lambda_i} = \phi(x)^T\left[\,\phi(x_1)\;\phi(x_2)\cdots\;\phi(x_n)\,\right] v_i/\sqrt{\lambda_i} \\ &= \left[\,\kappa(x,x_1)\;\kappa(x,x_2)\cdots\;\kappa(x,x_n)\,\right] v_i/\sqrt{\lambda_i},\end{aligned} \tag{7}$$

or in matrix notation $\Phi U = KV\Lambda^{-1/2}$ ($U$ is a matrix with $u_i$ in the columns, $V$ is a matrix with $v_i$ in the columns and $\Lambda^{-1/2}$ is a diagonal matrix with elements $1/\sqrt{\lambda_i}$), i.e., also the projections may be expressed in terms of the kernel function without using $\phi$ explicitly. If the mapping by $\phi$ is not column centered the variance of the projection must be adjusted, cf. [5,6].

Kernel PCA is a so-called memory-based method: from Equation 7 we see that if $x$ is a new data point that did not go into building the model, i.e., finding the eigenvectors and -values, we need the original data $x_1, x_2, \ldots, x_n$ as well as the eigenvectors and -values to find scores for the new observations. This is not the case for ordinary PCA where we do not need the training data to project new observations.

**Some Popular Kernels.** Popular choices for the kernel function are stationary kernels that depend on the vector difference $x_i - x_j$ only (they are therefore invariant under translation in feature space), $\kappa(x_i, x_j) = \kappa(x_i - x_j)$, and homogeneous kernels also known as radial basis functions (RBFs) that depend on the Euclidean distance between $x_i$ and $x_j$ only, $\kappa(x_i, x_j) = \kappa(\|x_i - x_j\|)$. Some of the most often used RBFs are ($h = \|x_i - x_j\|$)

- multiquadric: $\kappa(h) = (h^2 + h_0^2)^{1/2}$,
- inverse multiquadric: $\kappa(h) = (h^2 + h_0^2)^{-1/2}$,
- thin-plate spline: $\kappa(h) = h^2 \log(h/h_0)$, or
- Gaussian: $\kappa(h) = \exp(-\frac{1}{2}(h/h_0)^2)$,

where $h_0$ is a scale parameter to be chosen. Generally, $h_0$ should be chosen larger than a typical distance between samples and smaller than the size of the study area.

## 4 Maximum Autocorrelation Factor Analysis

In maximum autocorrelation factor (MAF) analysis we maximize the autocorrelation of linear combinations, $a^T x(r)$, of zero-mean original (spatial) variables, $x(r)$. $x(r)$ is a multivariate observation at location $r$ and $x(r + \Delta)$ is an observation of the same variables at location $r + \Delta$; $\Delta$ is a spatial displacement vector.

### 4.1 Primal Formulation

The autocovariance $R$ of a linear combination $a^T x(r)$ of zero-mean $x(r)$ is

$$R = \mathrm{Cov}\{a^T x(r), a^T x(r + \Delta)\} \tag{8}$$
$$= a^T \mathrm{Cov}\{x(r), x(r + \Delta)\} a \tag{9}$$
$$= a^T C_\Delta a \tag{10}$$

where $C_\Delta$ is the covariance between $x(r)$ and $x(r + \Delta)$. Assuming or imposing second order stationarity of $x(r)$, $C_\Delta$ is independent of location, $r$. Introduce the multivariate difference $x_\Delta(r) = x(r) - x(r + \Delta)$ with variance-covariance matrix $S_\Delta = 2\ S - (C_\Delta + C_\Delta^T)$ where $S$ is the variance-covariance matrix of $x$ defined in Section 3. Since

$$a^T C_\Delta a = (a^T C_\Delta a)^T \tag{11}$$
$$= a^T C_\Delta^T a \tag{12}$$
$$= a^T (C_\Delta + C_\Delta^T) a/2 \tag{13}$$

we obtain

$$R = a^T (S - S_\Delta/2) a. \tag{14}$$

To get the autocorrelation $\rho$ of the linear combination we divide the covariance by its variance $a^T S a$

$$\rho = 1 - \frac{1}{2} \frac{a^T S_\Delta a}{a^T S a} \tag{15}$$
$$= 1 - \frac{1}{2} \frac{a^T X_\Delta^T X_\Delta a}{a^T X^T X a} \tag{16}$$

where the $n$ by $p$ data matrix $X$ is defined in Section 3 and $X_\Delta$ is a similarly defined matrix for $x_\Delta$ with zero-mean columns. $C_\Delta$ above equals $X^T X_\Delta/(n-1)$. To maximize $\rho$ we must minimize the Rayleigh coefficient $a^T X_\Delta^T X_\Delta a/(a^T X^T X a)$ or maximize its inverse.

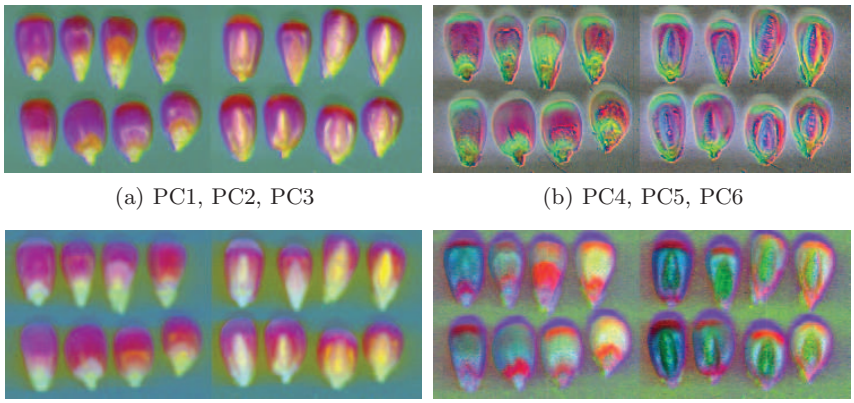
Unlike linear PCA, the result from linear MAF analysis is scale invariant: if $x_i$ is replaced by some matrix transformation $Tx_i$ corresponding to replacing $X$ by $XT$, the result is the same.

### 4.2 Kernel MAF

As with the principal component analysis we use the kernel trick to obtain an implicit non-linear mapping for the MAF transform. A detailed account of this is given in [10].

## 5 Results and Discussion

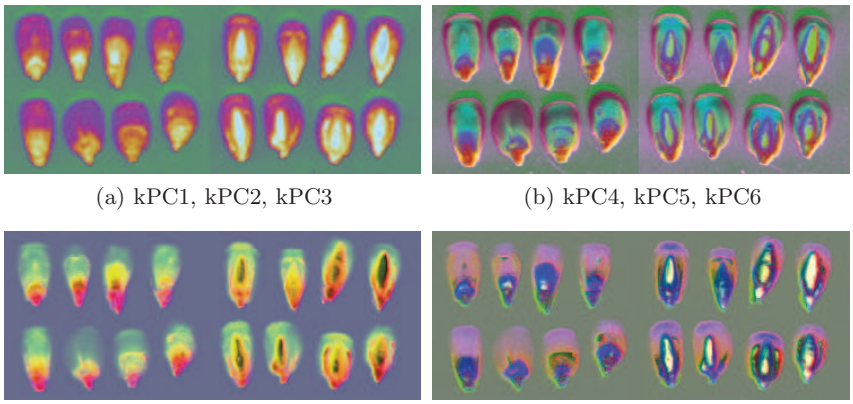
To be able to carry out kernel MAF and PCA on the large amounts of pixels present in the image data, we sub-sample the image and use a small portion termed the training data only. We typically use in the order $10^3$ training pixels (here ∼3,000) to find the eigenvectors onto which we then project the entire image termed the test data kernelized with the training data. A Gaussian kernel $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ with $\sigma$ equal to the mean distance between the training observations in feature space is used.



(a) PC1, PC2, PC3 (b) PC4, PC5, PC6

(c) MAF1, MAF2, MAF3 (d) MAF4, MAF5, MAF6

**Fig. 4.** Linear principal component projections of front and back sides of 8 maize kernels shown as RGB combination of factors (1,2,3) and (4,5,6) (two top panels), and corresponding linear maximum autocorrelation factor projections (bottom two panels)



(a) kPC1, kPC2, kPC3 (b) kPC4, kPC5, kPC6

(c) kMAF1, kMAF2, kMAF3 (d) kMAF4, kMAF5, kMAF6

**Fig. 5.** Non-linear kernel principal component projections of front and back sides of 8 maize kernel shown as RGB combination of factors (1,2,3) and (4,5,6) (two top panels), and corresponding non-linear kernel maximum autocorrelation factor projections (bottom two panels)

In Fig. 4 linear PCA and MAF components are shown as RGB combination of factors (1,2,3) and (4,5,6) are shown. The presented images are scaled linearly between $\pm 3$ standard deviations. The linear transforms both struggle with the background noise, local illumination and shadow effects, i.e., all these effects are enhanced in some of the first 6 factors. Also the linear methods fail in labeling the same kernel parts in same colors. On the other hand the kernel based factors shown in Fig. 5 have a significantly better ability to suppress background noise, illumination variation and shadow effect. In fact this is most pronounced in the kernel MAF projections. When comparing kernel PCA and kernel MAF the most striking difference is the ability of the kernel MAF transform to provide same color labeling of different maize kernel parts across all grains.

## 6 Conclusion

In this preliminary work on finding interesting projections of hyperspectral near infrared imagery of maize kernels we have demonstrated that non-linear kernel based techniques implementing kernel versions of principal component analysis and maximum autocorrelation factor analysis outperform the linear variants by their ability to suppress background noise, illumination and shadow effects. Moreover, the kernel maximum autocorrelation factors transform provides a superior projection in terms of labeling different maize kernels parts with same color.

## References


1. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosofical Magazine 2(3), 559–572 (1901)
2. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417–441, 498–520 (1933)
3. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, Heidelberg (2002)
4. Schölkopf, B., Smola, A., Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
5. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
6. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
7. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, Cambridge (2007)
8. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psykometrika 1, 211–218 (1936)
9. Johnson, R.M.: On a theorem stated by Eckart and Young. Psykometrika 28(3), 259–263 (1963)
10. Nielsen, A.A.: Kernel minimum noise fraction transformation (2008) (submitted)
11. Switzer, P.: Min/Max Autocorrelation factors for Multivariate Spatial Imagery. In: Billard, L. (ed.) Computer Science and Statistics, pp. 13–16 (1985)
12. Hoseney, R.C.: Principles of Cereal Science and Technology. American Association of Cereal Chemists (1994)
13. Belitz, H.-D., Grosch, W., Schieberle, P.: Food Chemistry, 3rd edn. Springer, Heidelberg (2004)

Appendix B

# Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images

Authors: Morten Arngren, Mikkel N. Schmidt, and Jan Larsen.

Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images. *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on, DOI 10.1109/MLSP.2009.5306262, p. 1-6, 2009.* Published.

# BAYESIAN NONNEGATIVE MATRIX FACTORIZATION WITH VOLUME PRIOR FOR UNMIXING OF HYPERSPECTRAL IMAGES

*Morten Arngren*[†°], *Mikkel N. Schmidt*[‡] *and Jan Larsen*[†]

[†] Technical University of Denmark, DTU Informatics, Richard Petersens Plads, DK-2800 Lyngby
[‡] University of Cambridge, Department of Engineering, Trumpington Street, Cambridge CB2 1PZ, UK
[°] FOSS Analytical A/S, Slangerupgade 69, DK-3400 Hilleroed

## ABSTRACT

In hyperspectral image analysis the objective is to unmix a set of acquired pixels into pure spectral signatures (endmembers) and corresponding fractional abundances. The Non-negative Matrix Factorization (NMF) methods have received a lot of attention for this unmixing process. Many of these NMF based unmixing algorithms are based on sparsity regularization encouraging pure spectral endmembers, but this is not optimal for certain applications, such as foods, where abundances are not sparse. The pixels will theoretically lie on a simplex and hence the endmembers can be estimated as the vertices of the smallest enclosing simplex. In this context we present a Bayesian framework employing a volume constraint for the NMF algorithm, where the posterior distribution is numerically sampled from using a Gibbs sampling procedure. We evaluate the method on synthetical and real hyperspectral data of wheat kernels.

## 1. INTRODUCTION

Traditional image acquisition and analysis is based on three color bands, red, green, and blue, which is sufficient for human visualization. In the context of identifying or extracting material constituents of e.g. foods, three channels are rarely enough.

Hyperspectral image analysis offers the opportunity to capture detailed spectral information for improved image analysis. Each of the observed pixels can be considered a mix of latent pure constituents. In foods, these constituents are typically *protein, starch,* $H_2O$, etc. The non-negative spectral signatures of these pure constituent are denoted *endmembers*. Each pixel also has a set of corresponding concentrations denoted *fractional abundances*, which are non-negative and must sum to one (additivity constraint). In a reflectance acquisition system the mixing can be considered linear according to Lambert-Beer's law. Non-linear mixing are usually neglected due to minor significance and increased modeling complexity [2].

Figure 1 illustrates a hyperspectral image of a wheat kernel with a corresponding pre-processed spectrum from $950 - 1650$nm.

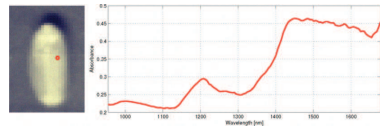


**Fig. 1**. Pseudo RGB image of a wheat kernel and the associated pre-processed spectrum for the selected pixel.

If we denote the $N$ observed $M$ dimensional spectra as the matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$, the linear mixing can be expressed as a $K$ rank linear factorization written as

$$\mathbf{X} = \mathbf{W}\mathbf{H} + \epsilon, \tag{1}$$

where the $K$ non-negative endmembers are contained in $\mathbf{W} \in \mathbb{R}_+^{M \times K}$, $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ holds the fractional abundances and $\epsilon$ denote the residual noise. The objective of spectral unmixing is to decomposition the pixels in order to extract the latent endmembers $\mathbf{W}$ and corresponding fractional abundances $\mathbf{H}$.

The non-negativity and sum-to-unity assumption of $\mathbf{H}$ implies a multidimensional simplex structure of the modeled data, where the vertices denote the endmembers. The observed data may not span the entire simplex due to the mixing of the constituents. For food applications pure endmembers appear rarely as the observed pixels are almost always a mix of constituents. This means the observed data will concentrate around the center of the simplex and very few samples can be expected at the vertices. The data acquisition is further subject to additive noise and thus the simplex structure will not be perfect. Figure 2 illustrates the simplex structure of wheat kernel image pixels. To unmix the hyperspectral image data, different approaches can be used. One approach is to analyse the data using convex geometrical methods. Plaza *et. al.* (2004) [9] gives a comparative analysis of a set of endmember extraction methods

978-1-4244-4948-4/09/$25.00 © 2009 IEEE

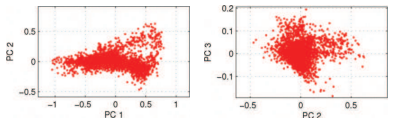



**Fig. 2**. Scatterplot of 1st, 2nd and 3rd principal component of the processed wheat kernel hyperspectral data.

based on geometrical analysis.

As the acquired data is non-negative by nature the Non-negative Matrix Factorization (NMF) has received much attention [8, 6, 7, 4]. The basic linear NMF model minimizes the squared error $||\mathbf{X} - \mathbf{WH}||^2$ subject to non-negativity constraint on the elements $\mathbf{W}$ and $\mathbf{H}$. This is however rarely sufficient to capture the endmembers in an unmixing problem and thus additional regularization is required. This can be expressed as part of the cost function as

$$C(\mathbf{W}, \mathbf{H}) = \tfrac{1}{2}||\mathbf{X} - \mathbf{WH}||^2 + \gamma J_w(\mathbf{W}) + \beta J_h(\mathbf{H}), \quad (2)$$

where $J_w(\mathbf{W})$ and $J_h(\mathbf{H})$ denote the regularization terms for the endmembers and fractional abundances respectively.

Sajda *et. al.* [10] develop an NMF algorithm with additional constraints on the amplitude of the estimated spectral components, with improved endmember determination. An $L_2$-norm sparsity prior on both the endmembers $\mathbf{W}$ and the fractional abundances $\mathbf{H}$ is incorporated by Pauca *et. al.* [8] also leading to improved estimation of the endmembers. A sparsity prior on the fractional abundances encourages pure spectra among the observed pixels, but this is rarely the case in foods and hence the sparsity prior is not useful.

Instead a prior can be based on the volume of the simplex. For a $K$-simplex in $M$-dimensional space for $K = M$ and with vertices $\{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_M\}$ the volume is given by $V_{simplex} = \frac{1}{K!}\det(\tilde{\mathbf{W}})$, where $\tilde{\mathbf{W}} = [\mathbf{w}_1 - \mathbf{w}_0\ \mathbf{w}_2 - \mathbf{w}_0 \ldots \mathbf{w}_M - \mathbf{w}_0]$ denote the vectors spanning the simplex. In our case $K << M$ and thus the simplex volume can be expressed as

$$V_{simplex} = \tfrac{1}{K!}\,|\det(\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})|^{1/2} \quad (3)$$

The simplex volume can be approximated by different means: Miao *et. al.* [3] present an NMF algorithm that incorporates the following volume regularization (denoted *mvc*)

$$J_w(\mathbf{W}) \propto \det{}^2\big(\mathbf{C} + \mathbf{B}\mathbf{U}_x^T(\mathbf{W} - \mu)\big), \quad (4)$$

where $\mathbf{U}_x$ is a projection matrix defined as the $K-1$ principal components of the observed data $\mathbf{X}$; $\mu$ denote the mean of $\mathbf{X}$; and $\mathbf{B}$ and $\mathbf{C}$ ensure a square matrix for the determinant. This volume regularization captures the volume of the centered simplex in the subspace defined by the $K-1$ eigenvectors. This approach is intuitive, but because of its dependency on the observed data it does not have an interpretation as a prior in a Bayesian framework.

A different volume regularization is proposed in Schachtner *et. al.* [11] based on the volume of the parallelepiped spanned by the endmembers. Based on this, the authors present a regularized NMF multiplicative update algorithm. The regularization is based on the determinant of the estimated endmembers expressed for the non-square case of $\mathbf{W}$ as (denoted *pp*)

$$J_w(\mathbf{W}) = \det(\mathbf{W}^T\mathbf{W}). \quad (5)$$

This regularization can be seen as a direct approximation of (3), where the absolute vectors $\mathbf{W}$ are used instead of the simplex spanning vectors $\tilde{\mathbf{W}}$. The omission of the square root leads to non-linear regularization, where large volumes will be penalized harder than small volumes. This regularization can be interpreted as a prior in a Bayesian framework, as it avoids the dependency to the observed data. It however suffers from including an offset mean $\mu_w$ leading to potential movement toward origo when minimizing the volume.

As opposed to regularized least squares, a different approach is to build a Bayesian framework for hyperspectral unmixing. This allows to define appropriate likelihood and prior distributions equivalent to the NMF formulation and to extract confidence intervals on a MAP estimate. In Ochs *et. al.* [5] a Bayesian Spectral Decomposition (BSD) algorithm is introduced based on an atomic prior. Moussaoui *et. al.* [4] present a Bayesian approach to separate non-negative mixtures of NIR data based on a hybrid Gibbs-Metropolis-Hastings sampling procedure. Schmidt *et. al.* [12] develop a Bayesian NMF based on a exponential sparsity prior and inverse gamma hyper priors.

In this paper we present a Bayesian NMF method with a volume prior for unmixing hyperspectral images. Our proposed method incorporates a determinant based prior inspired by the pp volume in (5). The additivity constraint on the fractional abundances is modeled as a uniform prior on the unit simplex. For inference in the model, a fast Gibbs sampling procedure is devised.

For comparison to our proposed Bayesian approach, the volume regularization terms mvc in (4) and pp in (5) is implemented in a projected gradient NMF framework. The additivity constraint is not included in the work of Schachtner *et. al.* [11] and is implemented as a soft-constraint by Miao *et. al.* [3] in their algorithm. This leads to a trade-off between describing the observed data and respecting the additivity constraint. In our projected gradient framework the additivity constraint is implemented with a variable substitution approach, which always fullfills the additivity constraint without sacrificing the description of the observed data. We evaluate the methods on artificial and real hyperspectral food image data.

In section 2 we present our Bayesian volume constrained NMF model as well as the Gibbs sampler. The synthetic and real data are described in section 3 with a short description of the initial pre-processing. Finally, in section 4 we evaluate how the proposed method can extract the true endmembers and corresponding fractional abundances, and compare to existing methods.

## 2. BAYESIAN NMF WITH VOLUME PRIOR

We here present a probabilistic Bayesian formulation of NMF with a prior that constrains the volume of the data simplex. We model the joint probability distribution of the endmembers, $\boldsymbol{W}$, and the fractional abundances, $\boldsymbol{H}$, as well as the noise, conditioned on the observed data, $\boldsymbol{X}$, and a set of model hyper-parameters, $\boldsymbol{\Theta}$. To this end we must choose a suitable noise model as well as reasonable prior distributions over all parameters in the model.

We model the noise, $\boldsymbol{\epsilon}$, as independent and identically distributed white Gaussian noise,which gives rise to the following likelihood function,

$$p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{\Theta}) = \prod_{n=1}^{N}\prod_{m=1}^{M}\mathcal{N}\left(x_{mn}|\boldsymbol{W}_{m:}\boldsymbol{H}_{:n},\sigma^2\right), \tag{6}$$

where $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x-\mu)^2}{-2\sigma^2}\right)$ is the Gaussian probability density function.

In our notation, matrices are denoted by capital bold letters. A matrix with two subscripts denotes the submatrix with the corresponding rows and columns, where a colon denotes all indices, and $\tilde{m}$ denotes all indices except $m$; a single element is denoted by $x_{mn}$.

The likelihood has a single parameter, the noise variance, $\sigma^2$, for which we choose an conjugate prior, i.e., an inverse-Gamma distribution,

$$p(\sigma^2) = \mathcal{IG}(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(-\frac{\beta}{\sigma^2}\right). \tag{7}$$

We choose a prior distribution for the endmembers, that encourages the simplex spanned by the estimated endmembers to be small. We base this prior on a determinant criterion, similar to [11], that measures the volume of the parallelepiped spanned by the endmembers. Furthermore, the prior includes the constraint that each element in the endmember matrix must be non-negative,

$$p(\boldsymbol{W}|\boldsymbol{\Theta}) \propto \begin{cases} \exp\left(-\gamma\det(\boldsymbol{W}^\top\boldsymbol{W})\right) & w_{mk}\geq 0 \\ 0 & \text{otherwise.}\end{cases} \tag{8}$$

The prior has a single parameter, $\gamma$, that determines the strength of the volume penalty. This parameter could possibly be modeled hierarchically; however, presently we set the parameter by hand.

For the fractional abundances, the prior must enforce non-negativity as well as the constraint that the abundances for each pixel must sum to unity. We choose a uniform prior on the unit simplex,

$$p(\boldsymbol{H}|\boldsymbol{\Theta}) \propto \begin{cases} 1 & h_{kn}\geq 0,\ \sum_{k=1}^{K}h_{kn}=1 \\ 0 & \text{otherwise.}\end{cases} \tag{9}$$

### 2.1. Gibbs sampler

Inference in the proposed probabilistic NMF model can be conducted using a Gibbs sampling procedure, in which we sequentially draw samples from the posterior conditional distribution of each parameter of the model. Due to our choice of priors, we can sample from all conditional distributions directly using standard methods, which obviates slow sampling procedures such as rejection sampling.

The conditional distribution of the noise variance is an inverse-Gamma,

$$p(\sigma^2|\boldsymbol{X},\boldsymbol{\theta}\backslash\sigma^2) = \mathcal{IG}(\sigma^2|\bar{\alpha},\bar{\beta}) \tag{10}$$

$$\begin{aligned}\bar{\alpha} &= \alpha + \frac{1}{2}NM - 1 \\ \bar{\beta} &= \beta + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(x_{mn}-\boldsymbol{W}_{m:}\boldsymbol{H}_{:n}\right)^2.\end{aligned} \tag{11}$$

The conditional distribution of the endmembers, which arises from the product of the Gaussian likelihood and the determinant-based prior, is a truncated Gaussian density,

$$p(w_{mk}|\boldsymbol{X},\boldsymbol{\theta}\backslash w_{mk}) \propto \begin{cases}\mathcal{N}(w_{mk}|\bar{\mu}_{mk},\bar{\sigma}^2_{mk}) & w_{mk}\geq 0 \\ 0 & \text{otherwise}\end{cases} \tag{12}$$

$$\bar{\sigma}^{-2}_{mk} = (\boldsymbol{H}_{k:}\boldsymbol{H}^\top_{k:})\sigma^{-2} + \gamma\big(D_{\tilde{k}\tilde{k}} - \boldsymbol{W}_{m\tilde{k}}\boldsymbol{A}_{\tilde{k}\tilde{k}}\boldsymbol{W}^\top_{m\tilde{k}}\big) \tag{13}$$

$$\bar{\mu}_{mk} = \bar{\sigma}^2_{mk}\Big(\big(\boldsymbol{W}_{m\tilde{k}}\boldsymbol{A}_{\tilde{k}\tilde{k}}\boldsymbol{W}^\top_{\tilde{m}\tilde{k}}\boldsymbol{W}_{\tilde{m}k}\big)\gamma + \big(\boldsymbol{X}_{m:}\boldsymbol{H}^\top_{k:} - \boldsymbol{W}_{m\tilde{k}}\boldsymbol{H}_{\tilde{k}:}\boldsymbol{H}^\top_{k:}\big)\sigma^{-2}\Big) \tag{14}$$

where $D_{\tilde{k}\tilde{k}} = \det(\boldsymbol{W}^\top_{:\tilde{k}}\boldsymbol{W}_{:\tilde{k}})$ and $\boldsymbol{A}_{\tilde{k}\tilde{k}} = \mathrm{adj}(\boldsymbol{W}^\top_{:\tilde{k}}\boldsymbol{W}_{:\tilde{k}})$. Samples from the truncated Gaussian distribution can be generated, e.g., using the method described by Geweke [1].

The conditional density of the fractional abundances is a constrained truncated Gaussian,

$$p(\boldsymbol{H}_{:n}|\boldsymbol{X},\boldsymbol{\theta}\backslash\boldsymbol{H}_{:n}) \propto \begin{cases}\mathcal{N}(\boldsymbol{H}_{:n}|\bar{\boldsymbol{\mu}}_n,\bar{\boldsymbol{\Sigma}}_n) & h_{kn}\geq 0,\ \sum_{k=1}^{K}h_{kn}=1 \\ 0 & \text{otherwise}\end{cases} \tag{15}$$

$$\bar{\boldsymbol{\mu}}_n = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X}_{:n}, \quad \bar{\boldsymbol{\Sigma}}_n = \sigma^2 (\boldsymbol{W}^\top \boldsymbol{W})^{-1}. \quad (16)$$

In the remainder of our paper we denote the Bayesian framework presented *BayesNMF-Vol*.

## 3. DATA ACQUISITION AND PRE-PROCESSING

The hyperspectral image data was acquired using a line-scan NIR camera from 900-1700nm in 165 bands leading to a data cube of $320 \times lines \times 165$. Prior to the image data being subjected to our unmixing algorithm a series of pre-processing steps were carried out. Initially, the raw image data was compensated for white reference and dark current background spectra in order to remove light source characteristics and sensor offset. Because the camera sensor has very poor signal-to-noise ratio from $900 - 950$nm and $1650 - 1700$nm, these ranges were removed. In order to suppress noise, each line was scanned twice and averaged. Afterwards, each spectra was converted to absorbance in order to obey Lambert-Beer's law.

For reference we acquired spectra of pure food constituents, *protein*, *starch* and *oil*, using the camera system and pre-processed as described. Figure 3 illustrates the spectra for these pure compounds.

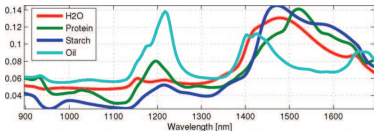



**Fig. 3**. Normalized spectra of pure basic food constituents. The water spectrum may suffer from poor SNR as $H_2O$ has extremely high absorbtion rates from $1400 - 1700$nm.

For our analysis, hyperspectral images of four wheat kernels were acquired on both front and back side leading to eight images. These images were pre-processed as described, and corrected for scatter effects. As sparse spectral peaks are not prominent in NIR data, low order scatter correction was applied as the residual from a first order polynomial fit. This approach preserves the simplex structure. The pure background pixels were then identified and removed from the data set, leaving a few around the kernel periphery. This segmentation was achieved by discriminating the first principal component applied on the reflectance data. Afterwards, the individual kernels were cropped. The final hyperspectral image data set was then represented as an unfolded matrix, $\mathbf{X} \in \mathbb{R}^{M \times N}$, where each $M$ dimensional column vector $\mathbf{x}$ is a pre-processed spectrum of a pixel. Figure 4 shows the pre-processed kernels prior to the background segmentation.

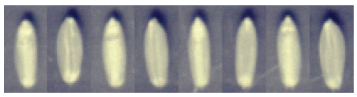


**Fig. 4**. Pseudo RGB image of wheat kernels 1-4 with pairwise front and backside.

## 4. SIMULATIONS

The two volume constrained NMF methods were evaluated in a set of simulations using synthetic data with different profiles and known labels. Afterwards our BayesNMF-Vol model was applied to real hyperspectral image data of wheat kernels.

### 4.1. Synthetic data

A synthetic dataset was produced by forming a 3-simplex in 3 dimensions based on randomly generated endmembers $\mathbf{W}$. Based on these endmembers, 100 samples were generated with full mixing profile and a noise variance $\sigma^2 = 10^{-3}$. This leads to a data matrix $\mathbf{X}$ of $3 \times 100$ illustrated in figure 5.

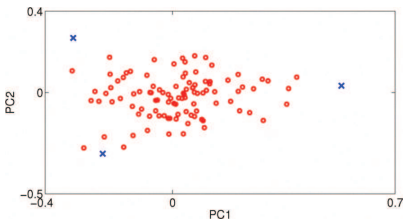



**Fig. 5**. Scatter plot of the synthetic data, 1st and 2nd PC. Circles denote datapoints and crosses denote endmembers.

Both the mvc, the pp regularized and the BayesNMF-Vol NMF algorithms were evaluated by applying the synthetic dataset to find their endmember MAP estimate $\mathbf{W}_{map}$ for different values of the regularization parameter $\gamma$. The posterior mode estimate was computed for the BayesNMF-Vol method. All algorithms were initialized from randomly selected observed data points. A small range of regularization parameter values $\gamma = \{0, .01, .02, .05, .1, .2, .5, 1\}$ were applied and the resulting scatter plots are shown in figure 6.

The plots depict the converged endmember estimates for all three approaches. The models successfully capture the data structure and increasing regularization encourages

smaller volumes. The advantage of our Bayesian model is that it allows us to estimate the full posterior producing confidence measures. The posterior mode estimates are further comparable with the performance of the two projected gradient NMF methods. In addition, the pp regularization indicate superior convergence properties, both in terms of significantly fewer iterations and slightly improved endmember estimate in average.

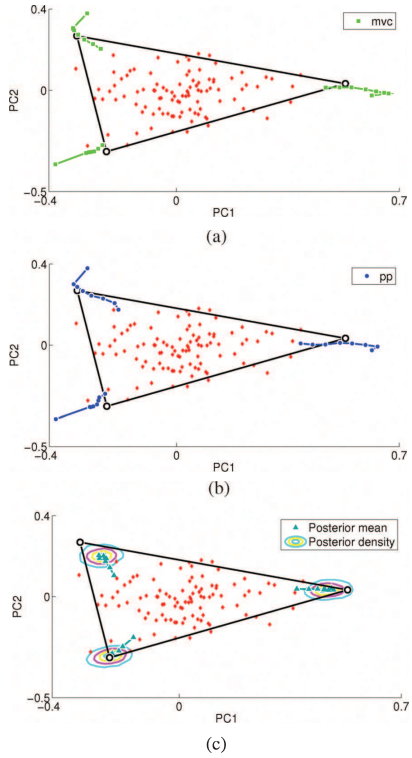



**Fig. 6**. Scatter plots of the observed data and the corresponding estimated endmembers for all 3 methods.

### 4.2. Wheat kernel data

A wheat kernel consists of many different constituents, where the majority is $H_2O$, *starch*, *protein* and *oil*, i.e. we expect to extract $4-5$ endmembers incl. background in our analysis. A set of reference concentrations for the compounds (typically protein) were unfortunately not available, so we are forced resort to subjective evaluation of the unmixing. The 4 kernels may also suffer from only small variations in the protein level between the pixel, i.e. strong correlation with other constituents. In such case the spectral profile of protein is almost impossible to extract. The corresponding PCA scatterplots of the wheat kernel pixels data indicate a simplex structure of 3-4 components (not shown).

The BayesNMF-Vol algorithm was initialized with pp regularized NMF endmember estimates in order to avoid long burn-in periods. A proper prior parameter was similarly found through empirical experiments. Figure 7 shows the estimated posterior mode endmembers and their spectral profiles for 10000 samples, $K = 3$ endmembers and $\gamma = 30$. A single decomposed wheat kernel is illustrated in figure 8.

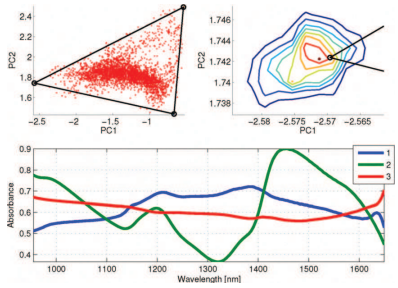



**Fig. 7**. Successful encapsulation of wheat kernel data. Note the burn-in samples on the right top illustration. The extracted spectral profiles, where the 2nd (green) is easily identified as starch comparing to figure 3.

Initially the background has been extracted by the model as the 3rd component. The starch spectral profile has also been identified very clearly along with a spatial distribution. Similarly the oil in the germ part can be identified primarily from the spatial distribution.

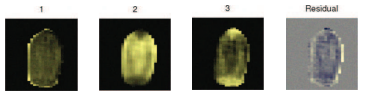



**Fig. 8**. Decomposition of backside of wheat kernel 3 (6th kernel in figure 4).

### 4.3. Discussion

All the volume based NMF methods described suffer from high sensitivity to the number of components, $K$. If K is too large, excessive components are typically split up in two.

The volume prior $p(\mathbf{W})$ based on (5) has a fatal sensitivity to linear dependencies among the estimated endmembers leading to a collapsing volume. This can occur if the estimated amount of components $K$ is higher than the true, where any excess endmember potentially fails to model the simplex. It can also occur for strong regularizations, i.e. large value of $\gamma$, leading to a collapse of the volume, as it is never allowed to expand and capture the data.

The experiments further reveal a high sensitivity to initializations as a bad start guess may require a temporary higher volume for rotation in order to settle correctly depending on data structure. If the volume prior $p(\mathbf{W})$ is set to penalize high volumes, then we might get stuck in a local minimum.

In addition our BayesNMF algorithm indicate a latent regularization on the estimated endmembers. For a flat prior distribution $p(\mathbf{W})$ with $\gamma = 0$, the BayesNMF-Vol still exhibit a small regularization encouraging small volumes. This means the observed data structure can not be encapsulated entirely. This effect is subject to further investigation.

The preliminary empirical analysis can easily be extended to more complex datasets in order to reveal a more detailed performance map in terms of large datasets with different mixing and noise profiles.

## 5. CONCLUSION

We have proposed a Bayesian NMF model employing a volume based prior more suitable for hyperspectral image analysis of foods. Using a synthetic data set our model indicate improved or similar unmixing performance compared to existing volume regulated NMF models.

Our BayesNMF-Vol model further extracts confidence intervals to the MAP estimate in terms of sampling variance.

Limitation to the performance of our model has also been identified and further research will focus on improving the framework.

## 6. REFERENCES


[1] J. Geweke, *Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities*, 1991 in Computing Science and Statistics: Proceedings of the 23rd Symposium, p.571-578

[2] Y. M. Masalmah, *Unsupervised Unmixing Of Hyperspectral Imagery Using The Constrained Positive Matrix Factorization*, Computing And Information Science And Engineering, University Of Puerto Rico, Mayaguez Campus, 2007

[3] L. Miao and H. Qi, *Endmember Extraction From Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization*, 2007 in Geoscience and Remote Sensing, IEEE Transactions on, Vol.45:3 p.765-777

[4] S. Moussaoui, D. Brie, A. Mohammad-Djafari and C. Carteret, *Separation of non-negative mixture of non-negative sources using a bayesian approach and MCMC sampling*, 2006 in IEEE Transactions on Signal Processing, Vol.54:11 p.4133-4145

[5] M. F. Ochs, R. S. Stoyanova, F. Arias-Mendoza and T. R. Brown, *A New Method for Spectral Decomposition Using a Bilinear Bayesian Approach*, 1999 in Journal of Magnetic Resonance, Vol.137:1 p.161-176

[6] P. Paatero and U. Tapper, *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, 1994 in Environmetrics, Vol.5:2 p.111-126

[7] L. Parra, C. Spence, P. Sajda, A. Ziehe and K. Mller, *Unmixing Hyperspectral Data*, 1999 in Neural Information Processing Systems, Vol.12

[8] P. V. Pauca, J. Piper and R. J. Plemmons, *Nonnegative Matrix Factorization for Spectral Data Analysis*, 2006 in Linear Algebra and Its Applications 416:1 p.29-47

[9] A. Plaza, P. Martnez, R. Prez and J. Plaza, *A Quantitative and Comparative Analysis of Endmember Extraction Algorithms From Hyperspectral Data*, 2004 in IEEE Transactions on Geoscience and Remote Sensing, Vol.42:3

[10] P. Sajda, S. Du and L. C. Parra, *Recovery of constituent spectra using non-negative matrix factorization*, 2003 in Proceedings of the SPIE - The International Society for Optical Engineering, Vol.5207:1 p.321-331

[11] R. Schachtner, G. Pppel, A. M. Tom and E. W. Lang, *Minimum Determinant Constraint for Non-negative Matrix Factorization*, 2009 in Lecture Notes in Computer Science, Vol.5441 p.106-113

[12] M. N. Schmidt, O. Winther and L. K. Hansen, *Bayesian non-negative matrix factorization*, 2009 in Lecture Notes in Computer Science (LNCS), Independent Component Analysis and Signal Separation, International Conference on, Vol.5441 p.540-547

Appendix C

# Unmixing of Hyperspectral Images using Bayesian Non-negative Matrix Factorization with Volume Prior

Authors: Morten Arngren, Mikkel N. Schmidt, and Jan Larsen.

Unmixing of Hyperspectral Images using Bayesian Non-negative Matrix Factorization with Volume Prior. *Journal of Signal Processing Systems, vol. 65 (3), p. 479-496, DOI: 10.1007/s11265-010-0533-2, 2011.* Published.

J Sign Process Syst (2011) 65:479–496
DOI 10.1007/s11265-010-0533-2

# Unmixing of Hyperspectral Images using Bayesian Non-negative Matrix Factorization with Volume Prior

**Morten Arngren · Mikkel N. Schmidt · Jan Larsen**

Received: 15 January 2010 / Revised: 31 May 2010 / Accepted: 13 September 2010 / Published online: 1 October 2010


**Abstract** Hyperspectral imaging can be used in assessing the quality of foods by decomposing the image into constituents such as protein, starch, and water. Observed data can be considered a mixture of underlying characteristic spectra (endmembers), and estimating the constituents and their abundances requires efficient algorithms for spectral unmixing. We present a Bayesian spectral unmixing algorithm employing a volume constraint and propose an inference procedure based on Gibbs sampling. We evaluate the method on synthetic and real hyperspectral data of wheat kernels. Results show that our method perform as good or better than existing volume constrained methods. Further, our method gives credible intervals for the endmembers and abundances, which allows us to asses the confidence of the results.

**Keywords** Bayesian source separation · Hyperspectral image analysis · Volume regularization · Gibbs sampling

M. Arngren (✉) · M. N. Schmidt · J. Larsen
DTU Informatics, Technical University of Denmark,
Bldg. 321, Richard Petersens Plads, 2800 Lyngby, Denmark
e-mail: ma@imm.dtu.dk

M. N. Schmidt
e-mail: mns@imm.dtu.dk

J. Larsen
e-mail: jl@imm.dtu.dk

M. Arngren
FOSS Analytical A/S, Slangerupgade 69,
3400 Hillerød, Denmark

## 1 Introduction

Classic image acquisition and analysis is based on three color bands (red, green, and blue), which is sufficient for human visualization. In the context of identifying or extracting material constituents of e.g. foods, these three wide channels are rarely enough. Hyperspectral image analysis can include more than 100 channels and hence provides the opportunity to capture detailed spectral information required in analysis of foods. Figure 1 illustrates a hyperspectral image of a wheat kernel with a corresponding pre-processed absorbance spectrum from 950–1,700 nm.

In a hyperspectral image acquisition system based on light transmission, mixing of the materials constituents can be considered linear according to Lambert-Beer's law. In a reflectance spectroscopy system, which we use in the experiments in this paper, non-linear mixing can occur, when incident light interacts with several constituent materials. For simplicity, however, it is reasonable to assume [16] that the mixing process is predominantly linear and that non-linear effects can be neglected.

Given a set of $N$ preprocessed $M$-dimensional mixed spectra, stored as columns in a matrix $\boldsymbol{X} \in \mathbb{R}_+^{M \times N}$, the unmixing problem consists of estimating the non-negative spectral signature of the pure constituent components (endmembers) as well as their relative contributions (fractional abundances) for each of the $N$ spectra. The linear mixing can then be expressed as a rank $K$ linear matrix factorization,

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{\epsilon}, \tag{1}$$

where $K$ is the number of endmembers. The non-negative spectral signatures of these endmembers are

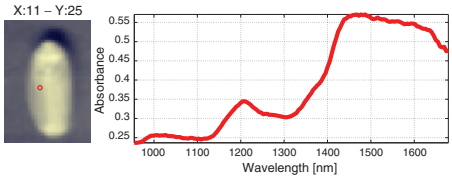



**Figure 1** Pseudo RGB image of a wheat kernel and the associated pre-processed spectrum for the selected pixel.

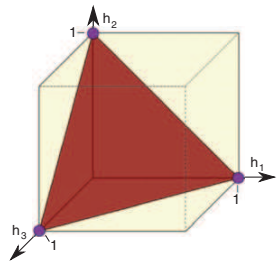



**Figure 2** Illustration of the unit 2-simplex in $\mathbb{R}^3$. The unit simplex is defined as $\{\boldsymbol{h} \in \mathbb{R}^M | \sum_{m=1}^{M} h_m = 1, \boldsymbol{h} \geq \boldsymbol{0}\}$. The abundance vectors (columns of $\boldsymbol{H}$) are confined to the unit simplex due to the non-negativity and sum-to-one constraints.

contained in the columns of $\boldsymbol{W} \in \mathbb{R}_+^{M \times K}$ and $\boldsymbol{H} \in \mathbb{R}_+^{K \times N}$ holds the fractional abundances for the $N$ elements. The matrix $\boldsymbol{\epsilon}$ denotes the residual noise. Each of the observed pixels can thus be considered a mix of latent pure constituents. In foods, these constituents are typically water, protein, starch, oil, etc.

In addition to the non-negativity constraint, the fractional abundances must sum to one[1] in order to maintain proper interpretation. The constraints imposed on the matrix factorization can thus be expressed as

$$w_{mk} \geq 0, \; h_{kn} \geq 0 \; \text{and} \; \sum_{k=1}^{K} h_{kn} = 1. \tag{2}$$

The non-negativity and sum-to-one assumption of the fractional abundances implies that the columns of $\boldsymbol{H}$ lie on a multidimensional simplex as illustrated in Fig. 2. The constraints and linear mixing further implies that the modelled data lie on a simplex in the observation space, where the endmembers are the vertices of the simplex. The simplex is illustrated for two endmembers in one dimension in Fig. 3a and for three endmembers in two dimensions in Fig. 3b.

The data might not span the entire simplex due to lack of mixing of the constituents. For food applications, pure endmembers appear rarely as the observed pixels are almost always a mix of constituents. This means the observed data will tend to concentrate around the center of the simplex and very few samples can be expected at the vertices. The data acquisition is further subject to additive noise and thus the simplex structure will not be perfect. Figure 4 illustrates the simplex structure of wheat kernel image pixels.

Different approaches can be used to unmix the hyperspectral image data. One approach is to analyse the data using convex geometrical methods. Plaza et al. [25] gives a comparative analysis of a set of endmember extraction methods based on geometrical analysis incl. the N-FINDR method [31], VCA [20], PPI [6] and other manually based endmember selection approaches. Most of these approaches rely on the assumption that pure pixels are present among the observed samples. This assumption might not be justified in hyperspectral images of food, and the method we propose does not require this assumption.

Another approach is based on statistical models of the data. As the acquired data is non-negative by nature, non-negative matrix factorization (NMF) has received wide attention [19, 22–24]. The basic linear NMF model minimizes the squared error $||\boldsymbol{X} - \boldsymbol{WH}||^2$ subject to non-negativity constraint on the elements in $\boldsymbol{W}$ and $\boldsymbol{H}$. These constraints are however rarely sufficient to capture the endmembers in an unmixing problem. Furthermore, the solution is not unique since any solution in which all the data are encapsulated by the endmembers will have the same cost; hence, additional regularization is required. This can be expressed through a regularized cost function,

$$C(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2}||\boldsymbol{X} - \boldsymbol{WH}||^2 + \gamma J_w(\boldsymbol{W}) + \beta J_h(\boldsymbol{H}), \tag{3}$$

where $J_w(\boldsymbol{W})$ and $J_h(\boldsymbol{H})$ are regularization terms for the endmembers and fractional abundances respectively.

Sajda et al. [26] present an NMF algorithm with additional constraints on the amplitude of the estimated spectral components, with improved endmember determination. An $\ell_2$-norm smoothness prior on both the

[1] In the literature, the constraints that abundances must sum to one is sometimes refered to as an *additivity constraint*.

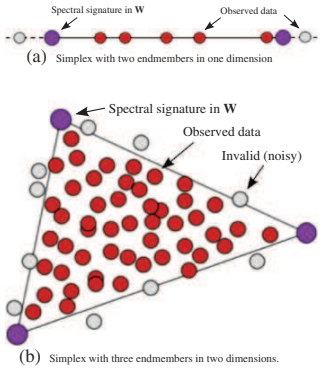



**Figure 3** Illustration of endmembers as the vertices of a simplex. The two illustrations show how the points violating the constraints are located outside the simplices formed by the endmembers (*purple circles*).

endmembers $\boldsymbol{W}$ and the fractional abundances $\boldsymbol{H}$ is incorporated by Pauca et al. [24], which also leads to improved estimation of the endmembers. Following Hoyer [15] an $\ell_1$-norm sparsity prior can also be incorporated in the NMF framework. A sparsity prior on the fractional abundances $\boldsymbol{H}$ encourages a pixel to be composed of only a few endmembers, but in analysis of food data this is rarely the case. Thus, the sparsity prior might not be useful in food applications. In this paper, we focus on volume regularization; however, if sparsity or smoothness is relevant, our method can easily be extended to include such priors.

A different approach is to incorporate a regularization based on the volume of the simplex spanned by the estimated endmembers. This encourages a decomposition in which the estimated endmembers lie closer to the data. The volume of a $(K-1)$-simplex with $K$ vertices $\{\boldsymbol{w}_{:1}, \ldots, \boldsymbol{w}_{:K}\}$[2] and $K < M$ is given by (denoted *vol*)

$$J_w^{\mathrm{vol}}(\boldsymbol{W}) = \frac{1}{K!}\sqrt{\det\left(\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}\right)}, \tag{4}$$

[2]In the notation used in this paper, matrices and vectors are denoted by capital and lower case bold letters respectively. Two subscripts denotes a sub matrix or sub vector with the corresponding rows and columns, where a colon denotes all indices, and $\bar{m}$ denotes all indices except $m$. For example, $\boldsymbol{w}_{:k}$ denotes the $k$th column of $\boldsymbol{W}$ and $\boldsymbol{w}_{m\bar{k}}$ denotes the $m$th row of $\boldsymbol{W}$ with the $k$th element removed. A single element of the matrix $\boldsymbol{W}$ is denoted by $w_{mk}$.

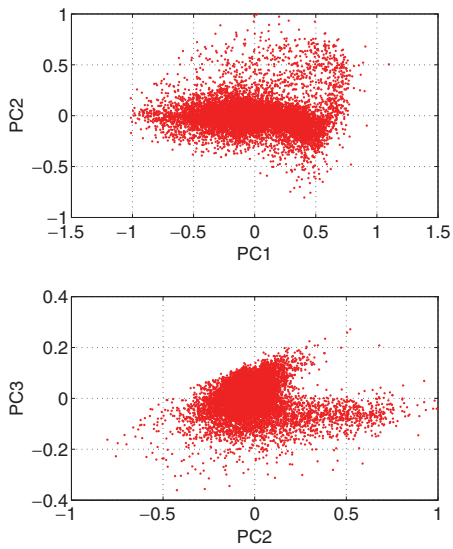



**Figure 4** Scatterplot of first and second principal component (PC) of the preprocessed wheat kernel hyperspectral data. The preprocessing pipeline used is described in Section 4.1.

where the column vectors of the matrix $\tilde{\boldsymbol{W}} = [\boldsymbol{w}_{:1} - \boldsymbol{w}_{:\rho} \ldots \boldsymbol{w}_{:\rho-1} - \boldsymbol{w}_{:\rho}, \boldsymbol{w}_{:\rho+1} - \boldsymbol{w}_{:\rho} \ldots \boldsymbol{w}_{:K} - \boldsymbol{w}_{:\rho}]$ point to the vertices of the simplex from an arbitrarily selected vertex, $\boldsymbol{w}_{\rho}$. For $K = M$ Eq. 4 reduces to $J_w^{\mathrm{vol}}(\boldsymbol{W}) = \frac{1}{K!}\det(\tilde{\boldsymbol{W}})$. The case of three endmembers in two dimensions is illustrated in Fig. 5.

Several authors have proposed algorithms for spectral unmixing that employ different volume based regularizations. Chan et al. [7] introduce the minimum-volume enclosing simplex (MVES) algorithm based on

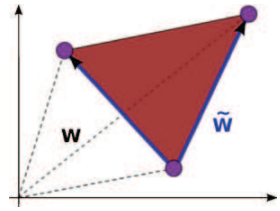



**Figure 5** The volume of a 2-simplex in two dimensions with $K = 3$ vertices, illustrating how the volume is computed based on relative endmember vectors.

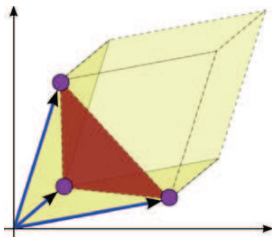


**Figure 6** The volume of a parallelepiped and the underlying 2-simplex in three dimensions.

convex analysis and a linear programming optimization scheme. Miao and Qi [18] present an NMF algorithm that incorporates the following volume regularization (denoted *minimum volume constraint* or *mvc*),

$$J_w^{\text{mvc}}(\boldsymbol{W}) \propto \det^2\left(\boldsymbol{C} + \boldsymbol{B}\boldsymbol{U}_x^\top\left(\boldsymbol{W} - \boldsymbol{\mu}\mathbf{1}^\top\right)\right), \tag{5}$$

where $\boldsymbol{U}_x$ is a projection matrix defined as the $K-1$ first principal components of the observed data $\boldsymbol{X}$, $\boldsymbol{\mu}$ denotes the sample mean of $\boldsymbol{X}$, $\boldsymbol{B} = \left[\begin{smallmatrix}\mathbf{1}^\top \\ \boldsymbol{O}\end{smallmatrix}\right]$ and $\boldsymbol{C} = \left[\begin{smallmatrix}\mathbf{0}^\top \\ \boldsymbol{I}\end{smallmatrix}\right]$, $\mathbf{1}$ and $\mathbf{0}$ are vectors of all ones and zeros respectively, $\boldsymbol{O}$ is an all zero matrix, and $\boldsymbol{I}$ is the identity matrix. This volume regularization captures the volume of the centered simplex in the subspace defined by the $K-1$ first principal eigenvectors and thus incorporates a noise reduction. This approach is quite intuitive, but because $\boldsymbol{U}_x$ depends on the observed data it cannot be given an interpretation as a prior in a Bayesian framework.

Schachtner et al. [27] propose a different volume regularization approach based on the squared volume of the parallelepiped spanned by the endmembers and origo, as shown in Fig. 6, and they propose an optimization method based on the NMF multiplicative update framework. The regularization term can be expressed as (denoted *parallelepiped* or *pp*)

$$J_w^{\text{pp}}(\boldsymbol{W}) = \det\left(\boldsymbol{W}^\top\boldsymbol{W}\right), \tag{6}$$

and can be seen as a surrogate to Eq. 4, where the absolute vectors (columns of $\boldsymbol{W}$) are used instead of the simplex spanning vectors (columns of $\tilde{\boldsymbol{W}}$). This measure, however, is sensitive to the location of the data simplex as opposed to the simplex volume in Eq. 4. This can potentially lead to movement toward origo, when minimizing the volume. Since the regularization is expressed in terms of a squared volume, large volumes will be penalized relatively stronger than small volumes.

Common for the three regularization terms based on the determinant is that they measure volume. Hence, they will tend to shrink the volume, but when the regularization is strong, the $K$-dimensional volume will collapse to a $(K-1)$-dimensional subspace of which the volume is zero leading to a regularization failure.

Another approach, which is not based on a determinant criterion, is to form an approximate volume regularization based on Euclidean distance measures. In the ICE algorithm, Berman et al. [4] implement a simplex volume measure as the sum of squared distances between all the vertices of the simplex. Equivalently, we incorporate a measure based on the sum of squared distances from the vertices to the centroid shown in Fig. 7 and given by (denoted *dist*),

$$J_w^{\text{dist}}(\boldsymbol{W}) = \sum_{m=1}^{M} \boldsymbol{w}_{m:}\left(\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)\boldsymbol{w}_{m:}^\top \tag{7}$$

$$= \sum_{k=1}^{K}\left\|\boldsymbol{w}_{:k} - \frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{:k'}\right\|_2^2. \tag{8}$$

This regularization term is not sensitive to the location of the simplex as the *pp* regularization is. With a large regularization, this measure will not collapse the simplex onto a lower dimensional subspace, but will shrink the simplex from each vertex towards the centroid. It further has the desirable property of being computationally inexpensive, as it does not require the computation of a determinant.

The *vol*, *pp*, and *dist* regularization terms can be given an interpretation as priors in a Bayesian framework, as they do not depend on the observed data. This is further discussed in the next section.

Common to the volume constrained spectral unmixing methods we have discussed so far is that they specify a regularized cost function and solve for the endmembers by numerical optimization. A different approach for hyperspectral unmixing is to build a probabilistic

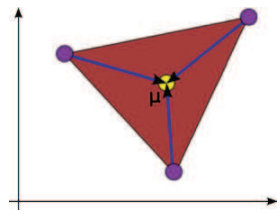



**Figure 7** The *dist* estimates the volume of a simplex based on the vertices Euclidean distance to the mean center.

J Sign Process Syst (2011) 65:479–496 

model and treat endmember extraction as a Bayesian inference problem. This requires the definition of an appropriate likelihood function and priors for the endmembers $\boldsymbol{W}$ and fractional abundances $\boldsymbol{H}$, which is closely related to the choice of cost function and regularization terms. The Bayesian approach further has the advantage of providing credible intervals in addition to an improved estimate of the endmembers and fractional abundances in the form of posterior mean estimates and thereby allowing for a more enhanced analysis.

Previous work on Bayesian spectral unmixing include Ochs et al. [21], who introduce a Bayesian spectral decomposition (BSD) algorithm based on an atomic prior. Schmidt and Laurberg [29] present a Bayesian NMF based on an exponential sparsity prior. Moussaoui et al. [19] present a sparse Bayesian method based on a hybrid Gibbs-Metropolis-Hastings sampling procedure for separating non-negative mixtures of NIR data, which are known to be sparse as opposed to the application considered in this paper. Dobigeon et al. [9] extend this approach to include a sum-to-one constraint. In related work Dobigeon et al. [8] propose a method for estimating endmembers in an appropriate subspace as opposed to the observation space.

In some applications, the number of endmembers can reasonably be assumed known in advance based on prior knowledge. For example, in the analysis of wheat kernels, knowledge of their biological properties can be used. To estimate the number of endmembers from data, different methods exist. A simple approach is to examine a PCA scatterplot to manually evaluate the required number of components to capture the simplex structure. Bioucas-Dias and Nascimento [5] present a fully automatic subspace based approach. The number of endmembers can also be estimated as part of a Bayesian framework. Schmidt et al. [30] describes an approach for Bayesian non-negative matrix factorization based on estimating the marginal likelihood (model evidence) and Eches et al. [11] present a method based on cross-dimensional sampling using reversible jump Markov chain Monte Carlo.

In this paper we extends our previous work [1] and present a Bayesian spectral unmixing method with a volume prior for unmixing hyperspectral images. The method embodies three different volume priors related to the *vol*, *pp*, and *dist* measures in Eqs. 4, 6 and 7. The method incorporates non-negativity constraints for the spectra as well as non-negativity and sum-to-one constraints for the fractional abundances. The number of endmembers are assumed to be known and fixed in advance. For model inference, we present a Markov chain Monte Carlo (MCMC) sampling procedure. The details of the method are described in the next section and we refer to the method as *BayesNMF-Vol*.

For comparison to our proposed Bayesian approach, we have implemented optimization based endmember extraction methods using the *mvc*, *pp*, and *dist* volume regularization terms in Eqs. 5, 6 and 7. Our implementation is similar to the methods presented in previous works, but is based on a projected gradient NMF framework (denoted *NMF-Vol*). The sum-to-one constraint for the fractional abundances is not included in the work of Schachtner et al. [27] and is implemented as a soft constraint by Miao and Qi [18] in their algorithm. This leads to a trade-off between describing the observed data and respecting the sum-to-one constraint. In our *NMF-Vol* framework the sum-to-one constraint is implemented using a variable substitution approach, which guarantees that the constraint is fulfilled. We evaluate both the *NMF-Vol* and *BayesNMF-Vol* methods on synthetic and real hyperspectral image data of foods. Both algorithms are available as a `Matlab` toolbox [3].

In Section 2 we present our Bayesian volume constrained NMF model as well as the MCMC sampling procedure. Simulations on synthetic data are described in Section 3, where we evaluate how the proposed method can extract the true endmembers and corresponding fractional abundances, and compare with existing methods. In Section 4 we describe a real wheat kernel data set and describe the pre-processing procedure used. We analyze the data using the proposed method and discuss the results. In Section 5 we discuss the general properties and limitations of our methods and conclude in Section 6.

## 2 Bayesian NMF with Volume Prior

The spectral unmixing methods presented in the previous section (except for *mvc*) can be given a probabilistic interpretation: They can be seen as maximum a posteriori (MAP) estimators. The data fit term $\frac{1}{2}||\boldsymbol{X} - \boldsymbol{WH}||^2$ corresponds to a Gaussian likelihood, and the regularization terms, $\gamma J_w(\boldsymbol{W})$ and $\beta J_h(\boldsymbol{H})$ correspond to priors over the parameters $\boldsymbol{W}$ and $\boldsymbol{H}$. Using a Bayesian approach, we are not limited to computing point estimates, such as the MAP estimator, but we can compute the full posterior distribution of the parameters. This can then be used to estimate the parameters and their credible intervals. When we are ultimately interested in a point estimate of the parameters, better estimators than the MAP can be computed, such as the posterior mean or median, which are optimal under squared and linear loss respectively.

### 2.1 Motivating the Bayesian Approach

To further motivate the use of Bayesian methods for spectral unmixing, we consider a very simple one-dimensional example, which clearly demonstrates the differences between the MAP and the Bayesian approach. Consider the case, where we have $N$ measurements, $x_n$,[3] which are points on the non-negative real line (see Fig. 8). We now wish to model these data using the non-negative linear factorization from Eq. 1, which in this case can be written as $x_n = w \cdot h_n + \epsilon_n$. For simplicity, we consider the noise free situation, $\epsilon_n = 0$, which results in a likelihood that requires all data points to be modelled exactly,

$$p(\boldsymbol{x}|w, \boldsymbol{h}) = \prod_{i=n}^{N} \delta\left(x_n - w \cdot h_n\right). \tag{9}$$

The prior over $w$ is chosen as a flat (improper) distribution over the non-negative real numbers, which can be thought of as a uniform distribution on the non-negative real line,

$$p(w) = \lim_{u_w \to \infty} \frac{1}{u_w} \mathbb{I}[0 \le w \le u_w] \tag{10}$$

$$\propto \mathbb{I}[w \ge 0]. \tag{11}$$

Here, $\mathbb{I}[\cdot]$ denotes an indicator function, which has the value one when its argument is true and zero otherwise. The prior for $\boldsymbol{h}$ is chosen as uniform between zero and one,

$$p(\boldsymbol{h}) = \prod_{n=1}^{N} \mathbb{I}\left[0 \le h_n \le 1\right]. \tag{12}$$

Our intuition about this model is that $w$ will take some value greater than the maximum data point, and $h_n$ will indicate the fractional distance at which $x_n$ lies. The prior on $w$ captures our ignorance about the location of $w$, and the prior on $h_n$ states that it corresponds to a proper fraction between zero and one. Data points can be generated from the model by first selecting a random $w$ from the prior,[4] and then for each data point selecting a random $h_n$ between zero and one. The generated data, $x_n$, will be uniformly distributed between 0 and $w$, and in the limit of infinitely many data points, intuition says that $w$ can consistently be estimated as the maximum of the observed data points.

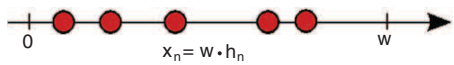



**Figure 8** One-dimensional linear factorization. Data $x_n$ (*circles*) are points on the non-negative real line. Since $h_n \in [0, 1]$, $w$ must be greater than the maximum data point.

Using Bayes' rule and collecting multiplicative constants, the posterior density of $w$ and $\boldsymbol{h}$ is given by

$$p(w, \boldsymbol{h}|\boldsymbol{x}) = \frac{1}{Z} \prod_{n=1}^{N} \delta(x_n - w \cdot h_n) \times \mathbb{I}[w \ge 0] \prod_{n=1}^{N} \mathbb{I}\left[0 \le h_n \le 1\right], \tag{13}$$

where $Z$ is a normalization constant. It is clear from the posterior, that a joint MAP estimate of $w$ and $\boldsymbol{h}$ can be found by choosing any $w$ greater than the maximum data value, and then choosing $h_n = x_n/w$. Any solution in which the range $[0, w]$ encapsulates the data, has the same posterior probability density, and the joint MAP estimate is thus not uniquely defined. For that reason, the joint MAP estimate is not particularly meaningful in this example; however, this is not because the model is ill defined. The posterior has a ridge of constant probability density, but in the Bayesian approach we are interested in probability mass rather than density. Insight can be gained by looking at the posterior marginals, which are found by integrating over the posterior density. This can be done analytically in this simple example, but requires more elaborate methods such as Markov chain Monte Carlo in the general multidimensional matrix factorization case.

The marginal density of $w$ is given by

$$p(w|\boldsymbol{x}) = \int_{\mathbb{R}^N} p(w, \boldsymbol{h}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{h} \tag{14}$$

$$\propto w^{-N} \cdot \mathbb{I}[x_{\max} \le w], \tag{15}$$

where $x_{\max} = \max_n(x_n)$. The posterior marginal captures the intuition that $w$ must be greater than the maximum data point, but moreover, it exhibits a polynomial decay, and thus its mass is concentrated in the region close to $x_{\max}$. In the limit $N \to \infty$, the marginal

[3] For simplicity, we use a slightly different notation in this section.

[4] In practice, of course, we cannot generate a random number from this improper prior, so we can think of generating $w$ from a uniform distribution between zero and $u_w$, where $u_w$ is some very large number.

J Sign Process Syst (2011) 65:479–496 

posterior will be infinitely peaked at $x_{\max}$. Similarly, the posterior marginal of $h_n$ is given by

$$p(h_n|\boldsymbol{x}) = \int_{\mathbb{R}^N} p(w, \boldsymbol{h}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{h}_{\bar{n}}\, \mathrm{d}w \tag{16}$$

$$\propto h_n^{N-2} \cdot \mathbb{I}\left[0 \leq h_n \leq \frac{x_n}{x_{\max}}\right], \tag{17}$$

which in the limit $N \to \infty$ is infinitely peaked at $\frac{x_n}{x_{\max}}$.

In the multidimensional hyperspectral unmixing problem discussed in this paper, data vectors are modelled as lying inside a $(K-1)$-simplex, and the objective is to identify the $K$ endmembers (the vertices of the simplex). With flat priors, the MAP estimate for this problem is not unique, as discussed previously, since any simplex which encapsulates the data vectors gives rise to a MAP solution, analogous to the simple example above. Using an more informative prior, such as the volume priors discussed previously, will encourage the simplex to be small, and thus overcome the problem of a non-unique MAP solution. As the example above suggests, when doing full Bayesian inference the non-uniqueness of the MAP solution is not an issue of concern—even when using flat priors. The reason is that although the maximum of the posterior is not unique, the posterior density itself is uniquely determined.

In addition to making the MAP estimate well determined, the different volume priors suggested in the literature also serve another purpose. Real data from hyperspectral imaging problems do not in general exactly obey the linear mixing property and thus the regularization parameter $\gamma$ in the volume prior can be used to push the algorithm towards a good solution. The data might also suffer from outlying data points and in such case the the regularization effect can similarly suppress the influence of the outliers in the simplex volume estimation and hence the determination of the model parameters. In the following we derive a Bayesian inference procedure for hyperspectral unmixing, which incorporates three different volume priors.

## 2.2 Model

We model the joint probability distribution of the endmembers, $\boldsymbol{W}$, and the fractional abundances, $\boldsymbol{H}$, as well as the noise, conditioned on the observed data, $\boldsymbol{X}$, and a set of model hyper-parameters, $\mathcal{H}$. To this end we choose a suitable noise model as well as reasonable prior distributions over all parameters in the model.

### 2.2.1 Noise Model

We model the noise, $\boldsymbol{\epsilon}$, as independent and identically distributed white Gaussian noise, which gives rise to the following likelihood function,

$$p\left(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \sigma^2\right) = \prod_{n=1}^{N}\prod_{m=1}^{M} \mathcal{N}\left(x_{mn}|\boldsymbol{w}_{m:}\boldsymbol{h}_{:n}, \sigma^2\right), \tag{18}$$

where $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{-2\sigma^2}\right)$ is the Gaussian probability density function. Note that the negative logarithm of the likelihood function corresponds to the squared error criterion which is the first term in Eq. 3.

The likelihood has a single parameter, the noise variance, $\sigma^2$, for which we choose a conjugate prior, i.e., an inverse-Gamma distribution,

$$p\left(\sigma^2|\alpha, \beta\right) = \mathcal{IG}\left(\sigma^2|\alpha, \beta\right) \tag{19}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\sigma^2}\right). \tag{20}$$

The hyperparameters, $\alpha$ and $\beta$, can be used to specify prior information about the noise variance. In our experiments, we take $\alpha$ and $\beta$ to the 0 limit, corresponding to a scale invariant improper prior.

### 2.2.2 Model of Fractional Abundances

For the fractional abundances, $\boldsymbol{H}$, the prior must enforce non-negativity as well as the constraint that the abundances for each pixel must sum to unity. We choose a uniform prior on the unit simplex, illustrated in Fig. 2,

$$p(\boldsymbol{H}) \propto \prod_{n=1}^{N} \mathbb{I}\left[\sum_{k=1}^{K} h_{kn} = 1\right] \prod_{k=1}^{K} \mathbb{I}\left[h_{kn} \geq 0\right], \tag{21}$$

which is arguably the most simple and least informative prior that expresses these constraints.

### 2.2.3 Model of Endmembers

We choose a prior distribution for the endmembers, that encourages the simplex spanned by the estimated endmembers to be small, and which includes the constraint that each element in the endmember matrix must be non-negative,

$$p(\boldsymbol{W}|\gamma) \propto e^{-\gamma J_w(\boldsymbol{W})} \prod_{m=1}^{M}\prod_{k=1}^{K} \mathbb{I}\left[w_{mk} \geq 0\right]. \tag{22}$$

The reason for choosing a prior proportional to $e^{-\gamma J_w(\boldsymbol{W})}$ is that the negative logarithm of the prior then corresponds to the regularization term, $\gamma J_w(\boldsymbol{W})$, in the cost function defined in Eq. 3. Thus, there is a direct parallel between the methods discussed in the introduction and our Bayesian probabilistic model. Furthermore, this prior gives rise to analytically tractable conditional densities required for the Gibbs sampler, provided that $J_w(\boldsymbol{W})$ is a quadratic form with respect to the components $w_{mk}$. Specifically, we consider three different volume measures,

$$J_w^{\mathrm{pp}}(\boldsymbol{W}) = \det\left(\boldsymbol{W}^\top \boldsymbol{W}\right), \tag{23}$$

$$J_w^{\mathrm{sv}}(\boldsymbol{W}) = \det\left(\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}\right) \text{ and} \tag{24}$$

$$J_w^{\mathrm{dist}}(\boldsymbol{W}) = \sum_{k=1}^{K} \left|\left| \boldsymbol{w}_{:k} - \frac{1}{K}\sum_{k'=1}^{K} \boldsymbol{w}_{:k'} \right|\right|_2^2. \tag{25}$$

$J_w^{\mathrm{pp}}(\boldsymbol{W})$ measures the squared volume of the parallelepiped defined by the endmembers and the origin, $J_w^{\mathrm{sv}}(\boldsymbol{W})$ is $(K!)^2$ times the squared volume of the simplex spanned by the relative endmember vectors, and $J_w^{\mathrm{dist}}(\boldsymbol{W})$ measures the sum of squared distances from the endmembers to their centroid. We have chosen these three measures of the simplex volume, because they can all be written as quadratic polynomials when considered as functions of a single element $w_{mk}$, which is easy to see, since the determinant is linear in its argument. For this reason, the prior has the form of a truncated Gaussian, when considered as a function of a single element $w_{mk}$. This makes it possible to derive a direct Gibbs sampling procedure, as we describe in the next section.

The parameter $\gamma$, which is common for these three volume priors, determines the strength of the volume penalty. In this paper, we study how the strength of volume prior influences the solutions and in our experiments we generate sets of solutions for a range of values of $\gamma$ as an explorative analysis. If a single solution is of interest, our approach could be extended to also estimate $\gamma$. For example, $\gamma$ could be be modelled hierarchically by choosing an appropriate prior distribution and including updates of $\gamma$ in the MCMC sampling procedure, or a good value for $\gamma$ can be found by cross validation or Bayesian model comparison.

*2.2.4 Posterior*

Using Bayes' rule, the posterior is given by

$$p(\boldsymbol{W}, \boldsymbol{H}|\boldsymbol{X}, \mathcal{H}) = \frac{p\left(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \sigma^2\right) p(\boldsymbol{H}) p(\boldsymbol{W}|\gamma) p\left(\sigma^2|\alpha, \beta\right)}{p(\boldsymbol{X})}, \tag{26}$$

where $\mathcal{H} = \{\alpha, \beta, \gamma\}$ are hyperparameters. In the following we let $\mathcal{P} = \{\boldsymbol{W}, \boldsymbol{H}, \sigma^2\}$ denote the parameters of the model.

2.3 Gibbs Sampler

A Gibbs sampling procedure [12, 13] can be used to infer the posterior distribution of the parameters of the model, $\mathcal{P}$. In Gibbs sampling, we sequentially draw samples from the posterior of each parameter, conditioned on all other parameters. It can be shown that the sequence of samples computed constitutes a Markov chain for which the stationary distribution is the posterior in which we are interested. Due to our choice of priors, we can sample from all conditional distributions directly using standard methods, which obviates slower sampling procedures such as rejection sampling.

*2.3.1 Sampling the Noise Variance*

Because we have chosen a conjugate prior for the noise variance, $\sigma^2$, its conditional distribution has the same functional form as the prior: An inverse-Gamma,

$$p\left(\sigma^2|\boldsymbol{X}, \mathcal{P}_{\setminus\sigma^2}\right) = \mathcal{IG}\left(\sigma^2|\bar{\alpha}, \bar{\beta}\right), \tag{27}$$

where the parameters are given by

$$\bar{\alpha} = \alpha + \frac{1}{2}NM, \tag{28}$$

$$\bar{\beta} = \beta + \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(x_{mn} - \boldsymbol{w}_{m:}\boldsymbol{h}_{:n}\right)^2. \tag{29}$$

Samples from this distribution can be generated using standard methods.

*2.3.2 Sampling Fractional Abundances*

The conditional density of the fractional abundances, $\boldsymbol{H}$, arises from the product of the Gaussian likelihood and the uniform prior on the unit simplex and it hence

J Sign Process Syst (2011) 65:479–496 

has the form of a Gaussian constrained to lie on the unit simplex,

$$p\left(\boldsymbol{h}_{:n}|\boldsymbol{X}, \mathcal{P}_{\backslash \boldsymbol{h}_{:n}}\right) \propto \mathcal{N}\left(\boldsymbol{h}_{:n}|\bar{\boldsymbol{\mu}}_n, \bar{\boldsymbol{\Sigma}}_n\right) \times \prod_{n=1}^{N} \mathbb{I}\left[\sum_{k=1}^{K} h_{kn} = 1\right] \prod_{k=1}^{K} \mathbb{I}\left[h_{kn} \geq 0\right], \tag{30}$$

and its posterior conditional parameters are given by

$$\bar{\boldsymbol{\mu}}_n = \left(\boldsymbol{W}^\top \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{x}_{:n}, \tag{31}$$

$$\bar{\boldsymbol{\Sigma}}_n = \sigma^2 \left(\boldsymbol{W}^\top \boldsymbol{W}\right)^{-1}. \tag{32}$$

Samples from the Gaussian density constrained to the unit simplex can be generated using the method of Dobigeon and Tourneret [10], or by the general procedure for sampling from a Gaussian subject to equality and inequality constraints described by Schmidt [28].

### 2.3.3 Sampling Endmembers

The conditional distribution of the endmembers, $\boldsymbol{W}$, arises from the product of the Gaussian likelihood and the volume penalizing prior. As noted earlier, the conditional prior has the form of a truncated Gaussian,

$$p\left(w_{mk}|\mathcal{P}_{\backslash w_{mk}}\right) \propto \mathcal{N}\left(w_{mk}|c_{mk}, s_{mk}^2\right) \mathbb{I}\left[w_{mk} \geq 0\right], \tag{33}$$

assuming $\boldsymbol{W}$ satisfies the non-negativity constraint. The values of $c_{mk}$ and $s_{mk}^2$ depend on the choice of $J_w(\boldsymbol{W})$. Using this, the conditional posterior is a truncated Gaussian

$$p\left(w_{mk}|\boldsymbol{X}, \mathcal{P}_{\backslash w_{mk}}\right) \propto \mathcal{N}\left(w_{mk}|\bar{\mu}_{mk}, \bar{\sigma}_{mk}^2\right) \mathbb{I}\left[w_{mk} \geq 0\right], \tag{34}$$

with parameters given by

$$\bar{\sigma}_{mk}^{-2} = s_{mk}^{-2} + \left(\boldsymbol{h}_{k:}\boldsymbol{h}_{k:}^\top\right)\sigma^{-2} \tag{35}$$

$$\bar{\mu}_{mk} = \bar{\sigma}_{mk}^2 \left(c_{mk} s_{mk}^{-2}, + \left(\boldsymbol{x}_{m:}\boldsymbol{h}_{k:}^\top - \boldsymbol{w}_{m\bar{k}}\boldsymbol{H}_{\bar{k}:}\boldsymbol{h}_{k:}^\top\right)\sigma^{-2}\right). \tag{36}$$

Samples from this distribution can be generated e.g. by efficient mixed rejection sampling [14, 17]. What is left is to derive expressions for $c_{mk}$ and $s_{mk}^2$ for each of the three volume measures in Eqs. 23–25.

For the determinant based measures, we use the following expression for the determinant of a symmetric matrix,

$$\det\begin{bmatrix} a & \boldsymbol{b}^\top \\ \boldsymbol{b} & \boldsymbol{C} \end{bmatrix} = a \det\left(\boldsymbol{C} - \frac{1}{a}\boldsymbol{b}\boldsymbol{b}^\top\right) \tag{37}$$

$$= a \det(\boldsymbol{C})\left(1 - \frac{1}{a}\boldsymbol{b}^\top \boldsymbol{C}^{-1}\boldsymbol{b}\right) \tag{38}$$

$$= a \det(\boldsymbol{C}) - \boldsymbol{b}^\top \mathrm{adj}(\boldsymbol{C})\boldsymbol{b}, \tag{39}$$

where Eq. 37 is the expression for the determinant of a block matrix, Eq. 38 follows from the matrix determinant lemma, and in Eq. 39 we have used the definition of the matrix adjugate. Using this for $-\gamma J_w^{\mathrm{pp}}$ we arrive at

$$s_{mk}^{-2} = \gamma\left(d_{\bar{k}\bar{k}} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{A}_{\bar{k}\bar{k}}\boldsymbol{w}_{m\bar{k}}^\top\right), \tag{40}$$

$$c_{mk} = s_{mk}^2 \gamma \boldsymbol{w}_{m\bar{k}}\boldsymbol{A}_{\bar{k}\bar{k}}\boldsymbol{W}_{\tilde{m}\bar{k}}^\top \boldsymbol{w}_{\tilde{m}k}, \tag{41}$$

where $d_{\bar{k}\bar{k}}$ and $\boldsymbol{A}_{\bar{k}\bar{k}}$ are the determinant and adjugate of $\boldsymbol{W}_{:\bar{k}}^\top \boldsymbol{W}_{:\bar{k}}$ respectively, and for $-\gamma J_w^{\mathrm{sv}}$ we get

$$s_{mk}^{-2} = \gamma\left(\tilde{d}_{\bar{k}\bar{k}} - \tilde{\boldsymbol{w}}_{m\bar{k}}\tilde{\boldsymbol{A}}_{\bar{k}\bar{k}}\tilde{\boldsymbol{w}}_{m\bar{k}}^\top\right), \tag{42}$$

$$c_{mk} = w_{m\rho} + s_{mk}^2 \gamma \tilde{\boldsymbol{w}}_{m\bar{k}}\tilde{\boldsymbol{A}}_{\bar{k}\bar{k}}\tilde{\boldsymbol{W}}_{\tilde{m}\bar{k}}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k}, \tag{43}$$

where $\rho \neq k$, and $\tilde{d}_{\bar{k}\bar{k}}$ and $\tilde{\boldsymbol{A}}_{\bar{k}\bar{k}}$ are the determinant and adjugate of $\tilde{\boldsymbol{W}}_{:\bar{k}}^\top \tilde{\boldsymbol{W}}_{:\bar{k}}$ respectively. For $-\gamma J_w^{\mathrm{dist}}$ we get

$$s_{mk}^{-2} = \gamma\left(\frac{K-1}{K}\right)^2, \quad c_{mk} = \frac{1}{K-1}\sum_{k' \neq k} w_{mk'}. \tag{44}$$

Details of the derivations can be found in an accompanying technical note [2].

## 3 Synthetic Data

Our two volume regularized spectral unmixing algorithms, *NMF-Vol* and *BayesNMF-Vol* were evaluated in a set of simulations using synthetic data.

### 3.1 Synthetic Data Generation

A set of synthetic datasets with different noise levels and dimensionality were generated with full mixing, i.e., the generated samples spanned the entire simplex. Several synthetic datasets were created with a different number of endmembers, varying data dimensionality,

**Table 1** List of synthetic data sets.

| ID | Endmembers $K$ | Dim. $M$ | Noise var. $\sigma^2$ | Samples $N$ |
|---|---|---|---|---|
| 1 | 3 | 3 | $10^{-4}$ | 3,000 |
| **2** | **3** | **3** | $\mathbf{10^{-3}}$ | **3,000** |
| **3** | **3** | **150** | $\mathbf{10^{-4}}$ | **3,000** |
| **4** | **3** | **150** | $\mathbf{10^{-2}}$ | **3,000** |
| 5 | 5 | 150 | $10^{-4}$ | 5,000 |
| 6 | 5 | 150 | $10^{-3}$ | 5,000 |
| **7** | **5** | **150** | $\mathbf{10^{-1}}$ | **5,000** |
| 8 | 5 | 150 | $10^{-4}$ | 50 |
| 9 | 5 | 150 | $10^{-3}$ | 50 |
| **10** | **5** | **150** | $\mathbf{10^{-2}}$ | **50** |

Boldfaced will be used for illustration.

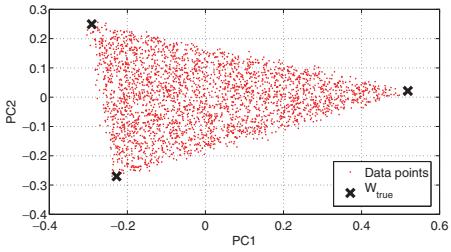



**Figure 9** Scatter plot of the synthetic data for, 1st and 2nd PC. Note the uniform distribution of the samples over the unity simplex.

and noise level. The first experiments had few endmembers in a low-dimensional space. The next included a realistic number of endmembers and data dimensions found in real hyperspectral images. Our final experiments were conducted with a small sample size, where the dimensionality of the data was greater than the number of samples. Table 1 lists the different synthetic data set profiles. Each dataset was based on uniform distributed randomly generated endmembers $\mathbf{W}$ used to produce different amounts of fractional abundance samples $\mathbf{H}$.

The samples were generated using the *simplex point picking* procedure leading to a uniform distribution of the generated samples $\mathbf{H}$ over the unit simplex. The procedure exploits the fact that the simplex can be considered a Dirichlet distribution with all parameters set to 1. Initially uniform random samples between [0; 1] are drawn as $h_{uni}$ and processed as $h = -\log(h_{uni})$. They are afterwards scaled to $h_{k,n} = h_{k,n} / \sum_{k'} h_{k',n}, \forall k$. Figure 9 illustrates an example of a synthetically generated dataset as a principal component scatterplot.[5]

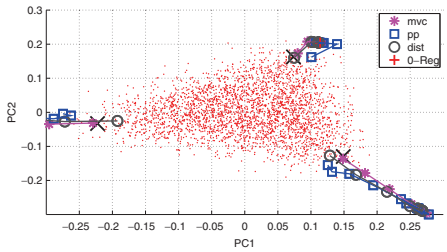
### 3.2 Simulations

The synthetic datasets were initially analyzed using *NMF-Vol* with three different regularizations; the *mvc* Eq. 5, *pp* Eq. 6 and *dist* Eq. 7 measures. All algorithms were initialized from randomly selected observed data points among $\mathbf{X}$ and a range of regularization parameter values $\gamma \in \{0, 10^{-6}, 10^{-5}, \ldots, 10^{4}\}$ were applied. Finally, the endmember MAP estimates $\mathbf{W}_{\text{map}}$ for the different regularizations were found. Similarly, our *BayesNMF-Vol* method was applied to the same datasets with a similar range of appropriate regularization. We used the tree different volume priors; *pp* Eq. 23, *sv* Eq. 24, and *dist* Eq. 25. For each regularization level we generated 3,000 Gibbs samples and disregarded the initial 2,000 samples as burn in to calculate the posterior mean endmember estimates. A more thorough simulation generating 12,000 Gibbs samples incl. 2,000 burn-in samples were used to produce and evaluate the credible intervals for $\gamma = 0$.

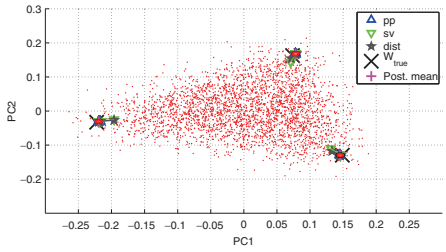
Results of a subset of our experiments (boldface in Table 1) are shown in Figs. 10, 11, 12, and 13 illustrating the converged endmember estimates for both algorithms with all regularization approaches and with interconnecting lines for the different regularization levels. For illustrative purposes, only results for a subset of the values of the regularization parameter $\gamma$ are shown, chosen such that the estimated endmembers lie in a range close to the true endmembers. The models successfully capture the data structures with more samples than dimensions and reveal similar performance in endmember extraction. The *BayesNMF-Vol* model successfully captures the endmembers on par with the three regular volume regularized NMF models. There is, however, a clear difference between the two approaches: With low regularization, the *NMF-Vol* methods find endmembers outside the data simplex, and thus need a suitable regularization to give a reasonable answer. This is especially visible in Fig. 10a. The *BayesNMF-Vol* with low regularization gives solutions close to the true endmembers, and higher regularization tends to shrink the simplex further as desired. A major advantage is the possibility to compute credible intervals for the endmembers, as depicted in Figs. 10c–13c. This means the confidence of the estimate can be evaluated. Increasing the regularization encourages smaller volumes, as expected, but is not necessarily required for the Bayesian method. We do expect,

[5] All scatterplots in this paper are presented as subspace projections onto the first and second principal component (PC) based on the data points.

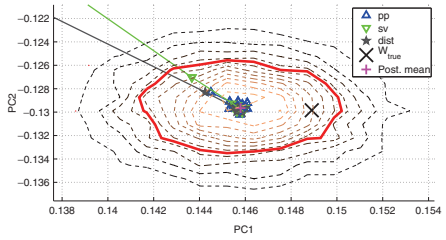
J Sign Process Syst (2011) 65:479–496 




(a) *NMF-Vol* showing MAP estimates.




(b) *BayesNMF-Vol* with posterior mean endmember estimates.




(c) *BayesNMF-Vol* zoomed with post. mean endmember estimates. Bold (red) line is 90% credible interval.

**Figure 10** Scatterplot of unmixing result with parameters $ID = 2$, $M = 3$, $K = 3$, $\sigma^2 = 10^{-3}$. Different regularization levels are shown with *interconnecting lines*. Outermost points correspond to no regularization, $\gamma = 0$.

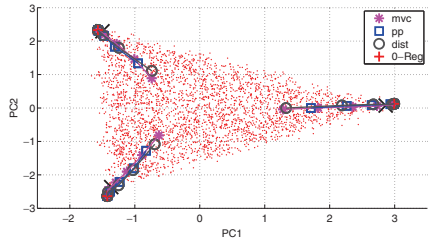
however, that it will be useful in real data to counteract noise and outliers.

In practice the different regularizations all suppress the noise to enhance the simplex structure in the data. Depending on the level of noise a suitable level of regularization can be found and hence the unmixing performance across the parameters are expected to be comparable.

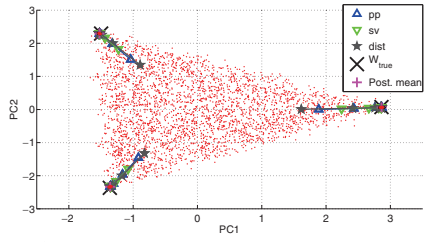
The data structures with fewer samples than dimension are a bit more difficult to unmix as seen in Fig. 13.

The estimated endmembers fluctuate more due to the lack of structural representation in the small amount of samples. This is also manifested in the larger variance for the BayesNMF-Vol method in Fig. 13c.

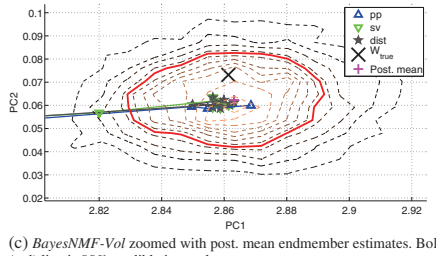
When the posterior distributions of endmembers are very narrow (as in the cases with many data points), we would expect them to be approximately isotropic. However, when the posterior distributions are wider (as in the case with few data points) we would not expect the posterior uncertainty to be equal in all dimensions—rather we would expect different uncertainty towards




(a) *NMF-Vol* showing MAP estimates.

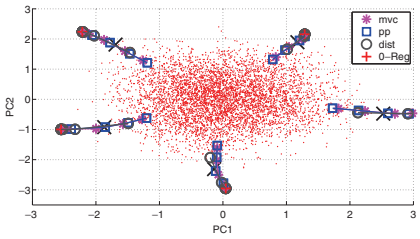



(b) *BayesNMF-Vol* with posterior mean endmember estimates.

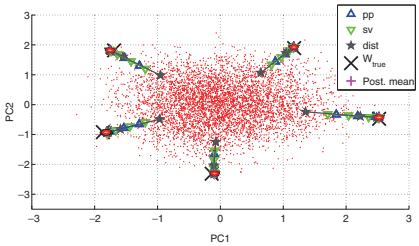



(c) *BayesNMF-Vol* zoomed with post. mean endmember estimates. Bold (red) line is 90% credible interval.

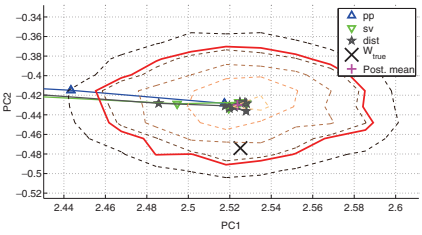
**Figure 11** Scatterplot of unmixing result with parameters $ID = 4$, $K = 3$, $M = 150$, $\sigma^2 = 10^{-2}$. Different regularization levels are shown with *interconnecting lines*. Outermost points correspond to no regularization, $\gamma = 0$.




(a) *NMF-Vol* showing MAP estimates.




(b) *BayesNMF-Vol* with posterior mean endmember estimates.




(c) *BayesNMF-Vol* zoomed with post. mean endmember estimates. Bold (red) line is 90% credible interval.

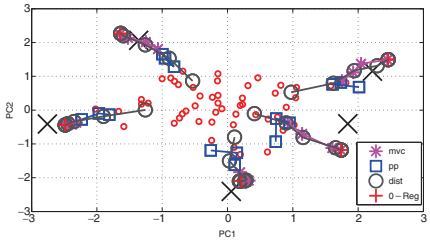
**Figure 12** Scatterplot of unmixing result with parameters $ID = 7$, $K = 5$, $M = 150$, $\sigma^2 = 10^{-1}$. Different regularization levels are shown with *interconnecting lines*. Outermost points correspond to no regularization, $\gamma = 0$.

the center vs. along the edge of the data point cloud. This is indeed also what we find and what the figures illustrate. Note, that the x and y axis in the figures are not identically scaled.

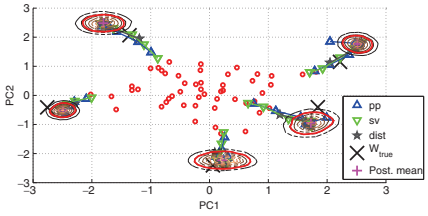
The differences between the determinant and distance based volume regularizations, as discussed in the introduction, is illustrated in Fig. 14. With increasing regularization strength, the *dist* prior shrinks toward the center of the simplex, whereas both the *pp* and the *sv* priors initially shrink the simplex and eventually collapses the simplex onto a line along the principal eigenvector. This result is intuitive, since the collapsed simplex has zero volume, but still extends along a direction that can explain the variance of the data.

## 4 Wheat Kernel Data

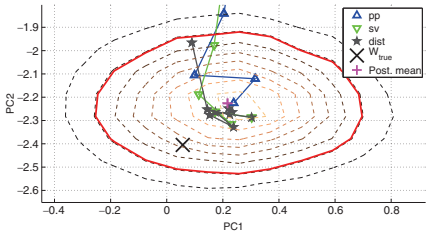
Our *BayesNMF-Vol* model was applied to real hyperspectral image data of wheat kernels.




(a) *NMF-Vol* showing MAP estimates.




(b) *Bayes NMF-Vol* with posterior mean endmember estimates.




(c) *Bayes NMF-Vol* zoomed with post. mean endmember estimates. Bold (red) line is 90 % credible interval.

**Figure 13** Scatterplot of unmixing result with parameters $ID = 10$, $K = 5$, $M = 150$, $\sigma^2 = 10^{-2}$. Different regularization levels are shown with *interconnecting lines*. Outermost points correspond to no regularization, $\gamma = 0$.

J Sign Process Syst (2011) 65:479–496 

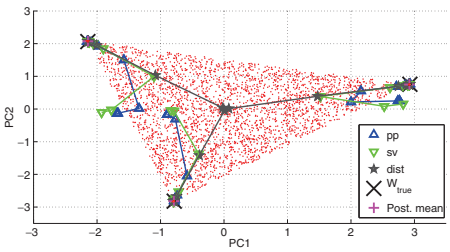



**Figure 14** Scatterplot of unmixing result with parameters $ID = 3$, $K = 3$, $M = 150$, $\sigma^2 = 10^{-4}$. For low regularization, the three volume priors give equal results. For large regularization the determinant based volume measures collapse onto a low-dimensional subspace (in this case, the area of the simplex collapses onto a line defined by the first PC) whereas the Euclidean based collapses onto a point (the simplex centroid).

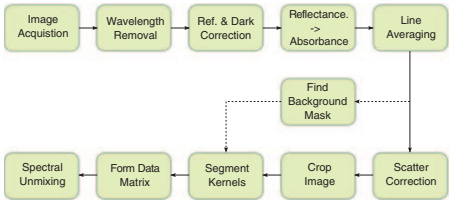



**Figure 16** Flowchart of pre-processing pipeline.

## 4.1 Data Acquisition and Pre-processing

A wheat kernel consists of many different constituents, where the majority is *water*, *starch*, *protein* and *oil*. Hence we expect to extract 4–5 endmembers including background in our analysis. For the acquisition of the image, a hyperspectral NIR line scan camera from Headwall Photonics Inc. is used from 900–1,700nm in 165 bands. Each acquired image is a 3-way tensor of size $320 \times L \times 165$ (where $L$ is the number of scanned lines). Two hyperspectral images of 14 wheat kernels were acquired on both front and back side, depicted in Fig. 15.

Prior to the image data being subjected to our unmixing algorithm a series of pre-processing steps were carried out shown in flowchart in Fig. 16. Initially the peripheral spectral range from 900–950 nm and 1,650–1,700 nm were removed due to the poor signal-to-noise ratio of the camera sensor. Afterwards, the raw image data was compensated for white reference and dark current background spectra to remove light source characteristics and sensor offset. In order to suppress noise, each line was scanned twice and averaged. The light scattering effects induced in the observed data were compensated in two steps. First each spectrum was converted to absorbance by a negative log-transform to obey Lambert-Beer's law for light transmission. Secondly a scatter correction step was used to align the spectra as shown in Fig. 17. As sparse spectral peaks are not prominent in NIR data, low order scatter correction was applied as the residual from a first order polynomial fit. This approach preserves the simplex structure as shown in Fig. 4.

For our analysis, the individual wheat kernels were extracted or cropped from the images by identifying and removing the pure background pixels from the data set. This segmentation was achieved by discriminating the first principal component applied on the image data depicted in Fig. 18. A few background pixels were left around the kernel periphery as illustrated in Fig. 19 in order to capture an entire grain kernel. This further allows the background to be identified

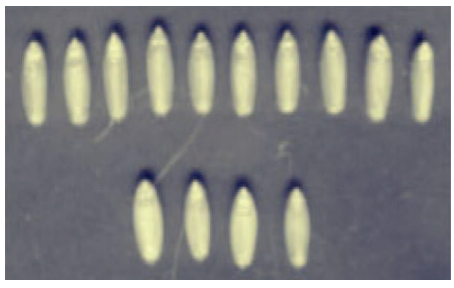


**Figure 15** The raw acquired image of the front side of the wheat kernels in pseudo colors.

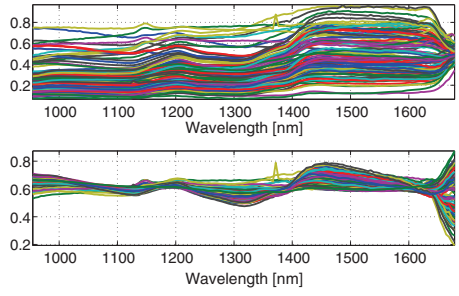



**Figure 17** The observed spectra before and after the affine scatter correction (above and below respectively).

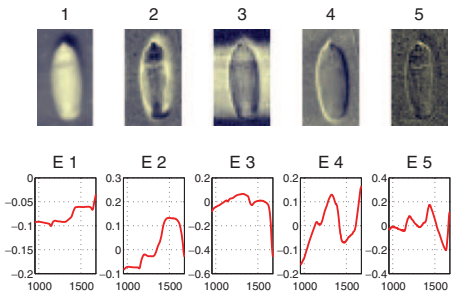
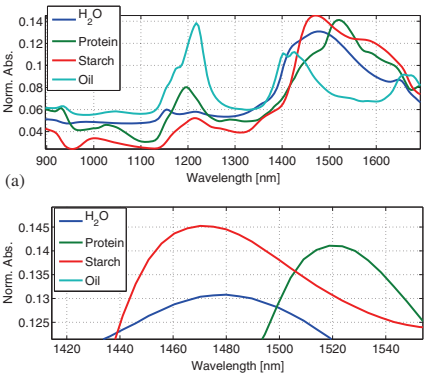



**Figure 18** Principal component images of the wheat kernel data set with corresponding associated PC spectra enumerated E1-E5.




(b) The protein and starch spectral profiles are quite similar, but exhibit a clear difference in peak location around 1500nm. This will prove useful when identifying estimated end members in food applications.

**Figure 20** Normalized spectra of pure basic food constituents. The water spectrum may suffer from poor SNR as $H_2O$ has extremely high absorption rates from 1,400–1,700 nm.

as a single endmember component in our subsequent unmixing. The final hyperspectral image data set was then represented as an unfolded matrix, $\boldsymbol{X} \in \mathbb{R}^{M \times N}$, where each $M$-dimensional column vector $\boldsymbol{x}_n$ is a pre-processed spectrum of a pixel.

For reference, spectra of pure food constituents, i.e., protein powder, starch powder and liquid oil, were acquired using the same camera system and pre-processed as described. These measurements act as references in evaluating unmixed spectra and are shown in Fig. 20. Initially the spectral profiles of protein and starch appear similar in most of the wavelength range. The most prominent spectral difference is around 1,450–1,550 nm as shown in Fig. 20b. Any relocation of the peak can be used to indicate different protein/starch mixtures and hence to interpret unmixing spectral results.

### 4.2 Simulations

The wheat kernel data comprises 14 grains, where 4 were selected with front and back side to be used for the unmixing, i.e., a total of 8 kernel images. Based on their biological properties, wheat kernels consists of many different constituents, where the majority is *water*, *starch*, *protein* and *oil*, i.e., we expect to extract 4–5 endmembers incl. background in our analysis.

As ground truth concentrations for the constituents are not available, we perform a subjective evaluation of the unmixing results. The 4 kernels may suffer from only small variations in the protein level between the pixels, which means that there is a strong correlation with other constituents. In such case the spectral profile of protein is almost impossible to extract. Principal component scatterplots of the wheat kernel data

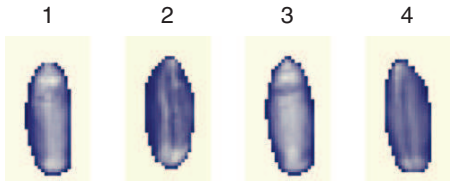



**Figure 19** A few segmented kernels with a little background included in the periphery.

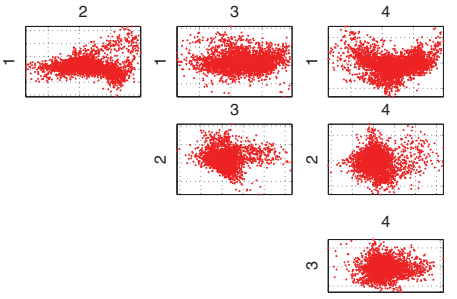


**Figure 21** PCA scatter plot of the wheat kernel data indicating a strong simplex structure in the first 4 PCs. This suggests 4 endmembers latent in the data set.

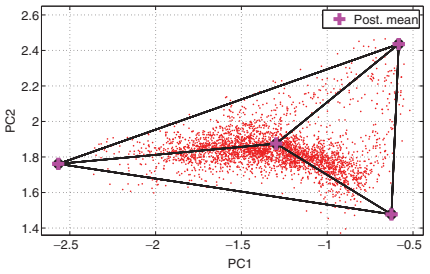
J Sign Process Syst (2011) 65:479–496 493

suggest a simplex structure of 4 components as illustrated in Fig. 21.

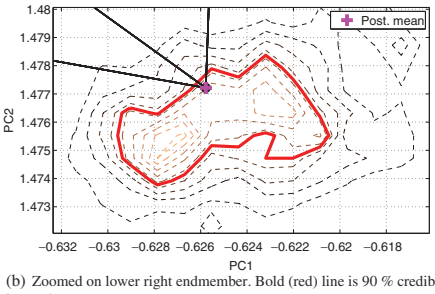
We analysed the data using the BayesNMF-Vol algorithm with *pp* volume prior. For initialization we used *pp*-regularized *NMF-Vol* endmember estimates in order to avoid unnecessary long burn-in periods. Analysis were conducted for a range of prior parameters $\gamma$, and results are shown for the value of $\gamma$ that gave the visually most clear unmixing. Figure 22 shows the estimated posterior mean endmembers for $K = 4$ endmembers and 10,000 Gibbs samples having disregarded the initial 10,000 burn-in samples.

The spectral profiles of the endmembers, illustrated in Fig. 23a, shows how the 2nd and 4th spectra designate the starch and protein content. The most prominent difference is the position of the peak at appr. 1,450 nm. indicating different mixtures of protein and starch as compared with the reference measurements in Fig. 20.

Of the 4 kernels (front and backside) a single decomposed wheat kernel is illustrated in Fig. 23b. Initially the background has been extracted by the model as the 1st component with the flat spectrum. The protein and starch spectral profile have also been identified very clearly along with a spatial distribution in the 2nd and 4th component. Similarly the oil in the germ part can be identified primarily from the spatial distribution. Finally, the residual reveals very little structure suggesting a successful decomposition.

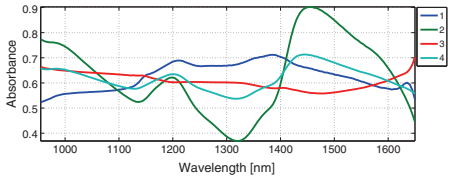



(a) Scatter plot of estimated simplex.

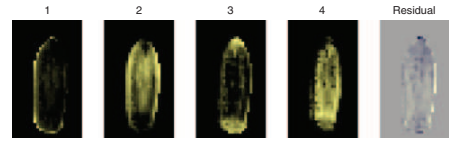



(b) Zoomed on lower right endmember. Bold (red) line is 90 % credible interval.

**Figure 22** BayesNMF-Vol analysis of wheat kernel data.




(a) Extracted spectral endmembers as the posterior mean.




(b) Image components of an unmixed wheat kernel as the posterior means.

**Figure 23** The 2nd and 4th extracted spectral endmember (*green* and *cyan*) are easily identified as starch/protein matrix comparing to Fig. 20.

## 5 Discussion

All the volume based NMF methods described show light sensitivity to the number of components, $K$. If $K$ is too large, excessive components are typically split up in two. In contrast, components will be merged if $K$ is too small. We have found that the regularization terms/priors based on the volume as computed using the determinant sometimes have a fatal sensitivity to linear dependencies among the estimated endmembers leading to a collapsing volume. This can occur if the number of components $K$ used in the analysis is greater than the true number of components in the data. In such situation, the excess endmember can cause the simplex to collapse and the volume regularization will be rendered inoperative. The Euclidean distance based volumes does not suffer from the same issue as it simply shrinks the volume according to the strength of the regularization/prior. It can also occur that a strong regularizations, i.e., large value of $\gamma$, leads to a collapse of the volume, as it is never allowed to expand and capture the data.

Our *BayesNMF-Vol* algorithm conducts light regularization in it self on the estimated endmembers as discussed. For a flat prior distribution $p(\mathbf{W})$ with $\gamma = 0$, the BayesNMF-Vol still encourages small volumes and does not necessarily require regularization to give meaningful results. This also means the observed data point may not be encapsulated entirely when using *BayesNMF-Vol* with $\gamma = 0$.

The computational performance of the *NMF-Vol* and *BayesNMF-Vol* are obvious very different due to the sampling nature of the latter. In our experiments approximately 5–7 iterations in the *NMF-Vol* algorithm correspond to 1 Gibbs sample in our *BayesNMF-Vol* method depending on model parameters. The MAP estimate is typically found is less than 3,000 iteration for the *NMF-Vol* algorithm depending on the structure of the data, noise levels, initialization etc. Assuming 20,000 Gibbs samples are sufficient to achieve a reasonable estimation of the credible intervals (incl. burn-in period), our *BayesNMF-Vol* method is then ∼30–50 times slower than the *NMF-Vol* algorithm, but has the advantage of providing a posterior mean estimate instead of a MAP estimate.

The synthetic data we used were limited to full mixing profiles, where the entire simplex is spanned with generated samples. In many real world datasets this might not always be true, as for our wheat kernel dataset for instance. Future improvement would therefore be to generate synthetic dataset with different mixing profiles to evaluate our methods performance in this area.

Both the *NMF-Vol* and *BayesNMF-Vol* are available as a `Matlab` toolbox [3] with a few examples.

## 6 Conclusion

We have proposed a Bayesian method for spectral unmixing, employing a volume based prior suitable for hyperspectral image analysis of foods. Results on synthetic data sets indicate similar or better unmixing performance compared to existing volume regulated NMF models and can further give predictive distributions, posterior means and credible intervals.

Our volume based priors has proven successful in an unmixing framework as opposed to existing methods employing sparsity based priors, which are not relevant in food applications due to the mixing nature of food constituents.

The limitation and properties of our *BayesNMF-Vol* method has been identified in particular in terms of the volume measures incorporated. The Euclidean volume measure proved faster and equally good in performance as the determinant based metrics.

In a concrete food application of decomposing wheat kernels into constituents our methods prove successful and can be used as part of assessment of the quality of foods.

## References


1. Arngren, M., Schmidt, M., & Larsen, J. (2009). Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images. In *Machine learning for signal processing, IEEE workshop on (MLSP)*. doi:10.1109/MLSP.2009.5306262.
2. Arngren, M., Schmidt, M. N., & Larsen, J. (2010). *Unmixing of hyperspectral images using Bayesian nonnegative matrix factorization with volume prior*. Technical note. http://www.imm.dtu.dk/pubdb/p.php?5836.
3. Arngren, M., Schmidt, M. N., & Larsen, J. (2010). *Unmixing of hyperspectral images using bayesian nonnegative matrix factorization with volume prior,* `Matlab` *toolbox*. http://www.imm.dtu.dk/pubdb/p.php?5834.
4. Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., & Huntington, J. (2004). ICE: A statistical approach to identifying endmembers in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing, 42*(10), 2085–2095.
5. Bioucas-Dias, J., & Nascimento, J. (2008). Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing, 46*(8), 2435–2445.
6. Boardman, J. W., Kruse, F. A., & Green, R. O. (1995). Mapping target signatures via partial unmixing of aviris data. In *Summaries of JPL airborne earth science workshop*.
7. Chan, T. H., Chi, C. Y., Huang, Y. M., & Ma, W. K. (2009). A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing, 57*(11), 4418–4432.
8. Dobigeon, N., Moussaoui, S., Coulon, M., Tourneret, J. Y., & Hero, A. O. (2009). Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery. *IEEE Transactions on Signal Processings, 57*(11), 4355–4368.
9. Dobigeon, N., Moussaoui, S., Tourneret, J. Y. Y., & Carteret, C. (2009). Bayesian separation of spectral sources under nonnegativity and full additivity constraints. *Signal Processing, 89*(12), 2657–2669.
10. Dobigeon, N., & Tourneret, J. Y. (2007). *Efficient sampling according to a multivariate Gaussian distribution truncated on a simplex*. Tech. Rep., IRIT/ENSEEIHT/TeSA.
11. Eches, O., Dobigeon, N., & Tourneret, J. Y. (2010). Estimating the number of endmembers in hyperspectral images using the normal compositional model and a hierarchical Bayesian algorithm. *IEEE Journal of Selected Topics in Signal Processing, 4*(3), 582–591.
12. Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*(410), 398–401.
13. Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741. doi:10.1109/TPAMI.1984.4767596.
14. Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear

constraints and the evaluation of constraint probabilities. In *Proceedings the 23rd symposium on the interface between computer sciences and statistics* (pp. 571–578). doi:10.1.1.26.6892.
15. Hoyer, P. O. (2002). Non-negative sparse coding. *Neural Networks for Signal Processing, 7*, 557–565.
16. Masalmah, Y. M. (2007). *Unsupervised unmixing of hyperspectral imagery using the constrained positive matrix factorization*. Ph.D. thesis, Computing And Information Science And Engineering, University Of Puerto Rico, Mayagez Campus.
17. Mazet, V., Brie, D., & Idier, J. (2005). Simulation of postive normal variables using several proposal distributions. In *IEEE workshop on statistical signal processing* (pp. 37–42). doi:10.1109/SSP.2005.1628560.
18. Miao, L., & Qi, H. (2007). Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing, 45*(3), 765–777.
19. Moussaoui, S., Brie, D., Mohammad-Djafari, A., & Carteret, C. (2006). Separation of non-negative mixture of non-negative sources using a Bayesian approach and mcmc sampling. *IEEE Transactions on Signal Processing, 54*(11), 4133–4145.
20. Nascimento, J. M. P., & Dias, J. M. B. (2005). Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing, 43*(4), 898–910.
21. Ochs, M. F., Stoyanova, R. S., Arias-Mendoza, F., & Brown, T. R. (1999). A new method for spectral decomposition using a bilinear bayesian approach. *Journal of Magnetic Resonance, 137*(1), 161–176.
22. Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetricsm 5*(2), 111–126.
23. Parra, L., Spence, C., Sajda, P., Ziehe, A., & Mller, K. R. (1999). Unmixing hyperspectral data. *Neural Information Processing Systems, 12*, 942–948.
24. Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications, 416*(1), 29–47.
25. Plaza, A., Martnez, P., Prez, R., & Plaza, J. (2004). A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing, 42*(3), 650–663.
26. Sajda, P., Du, S., & Parra, L. (2003). Recovery of constituent spectra using non-negative matrix factorization. *Proceedings of the SPIE - The International Society for Optical Engineering, 5207*(1), 321–331.
27. Schachtner, R., Pppel, G., Tom, A. M., & Lang, E. W. (2009). Minimum determinant constraint for non-negative matrix factorization. *Lecture Notes in Computer Science, 5441/2009*, 106–113.
28. Schmidt, M. N. (2009). Linearly constrained bayesian matrix factorization for blind source separation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1624–1632).
29. Schmidt, M. N., & Laurberg, H. (2008). Nonnegative matrix factorization with Gaussian process priors. *Computational Intelligence and Neuroscience* p. Article ID 361705.
30. Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *Independent component analysis and signal separation, international conference on lecture notes in computer science (LNCS)* (Vol. 5441, pp. 540–547).
31. Winter, M. E. (1999). N-findr: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. *Proceedings of SPIE - The International Society for Optical Engineering, 3753*, 266–275.

**Morten Arngren** received his B.Sc.E.E. degree from Copenhagen University College of Engineering in 1997 and a M.Sc. degree in applied math from the Technical University of Denmark in 2007. He has more than 10 years of experience in the software industry and is currently an Industrial PhD. Student at FOSS Analytical A/S in Denmark. His primary research include machine learning and hyperspectral image analysis for food applications.

**Mikkel N. Schmidt** received the B.Sc. degree in electronic engineering from Herning Institute of Business Administration and Technology (2001), the M.Sc. degree in engineering from Aalborg University (2003), and the Ph.D. degree in mathematical modeling from the Technical University of Denmark (2008). He has worked as a postdoctoral researcher at Cambridge University (2008–2009) and is now a researcher at the Technical University of Denmark in the Cognitive Systems group. His main research interest is Bayesian statistical modeling.

**Jan Larsen** received the M.Sc. and Ph.D. degrees in electrical engineering from the Technical University of Denmark (DTU) in 1989 and 1994.

Dr. Larsen is Associate Professor of Digital Signal Processing at Department of Informatics and Mathematical Modelling, DTU. Jan Larsen has authored and co-authored more than 100 papers and book chapters within the areas of nonlinear statistical signal processing, machine learning, neural networks and datamining with applications to biomedicine, monitoring systems, multimedia, and webmining. He has participated in more than ten national and international research programs, and has served as reviewer for many international journals, conferences, publishing companies and research funding organizations. As regards synergistic activities he took part in conference organizations, among others the IEEE Workshop on Machine Learning for Signal Processing (formerly Neural Networks for Signal Processing) 1999–2009. He is Director of Network for Danish Sound Technology (2009–2012), past chair of the IEEE Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2005–2007), and chair of IEEE Denmark Section's Signal Processing Chapter (2002-). He is a senior member of The Institute of Electrical and Electronics Engineers. Other professional committee participation includes: Member of the Technical Committee 14: Signal Analysis for Machine Intelligence of the International Association for Pattern Recognition, 2006-; Steering committee member of the Audio Signal Processing Network in Denmark, 2006-. Editorial Board Member of Signal Processing, Elsevier, 2006–2007; and guest editorships involves IEEE Transactions on Neural Networks; Journal of VLSI Signal Processing Systems; and Neurocomputing.

For a complete CV, see http://www.imm.dtu.dk/~jl.

Appendix D

# Unmixing of Hyperspectral Images using Bayesian Non-negative Matrix Factorization with Volume Prior, Supplementary material.

Authors: Morten Arngren, Mikkel N. Schmidt, and Jan Larsen.

Supplementary material for Unmixing of Hyperspectral Images using Bayesian Nonnegative Matrix Factorization with Volume Prior.
http://www2.imm.dtu.dk/pubdb/p.php?5836, 2011, Published.

# Unmixing of Hyperspectral Images using Bayesian Nonnegative Matrix Factorization with Volume Prior

TECHNICAL NOTE v1.0

**Morten Arngren**
Technical University of Denmark, DTU Informatics, Bldg. 321, Richard Petersens Plads, DK-2800 Lyngby
and FOSS Analytical A/S, Slangerupgade 69, DK-3400 Hillerød
ma@imm.dtu.dk,moa@foss.dk, info@arngren.com

**Mikkel N. Schmidt**
Technical University of Denmark, DTU Informatics, Bldg. 321, Richard Petersens Plads, DK-2800 Lyngby
mns@imm.dtu.dk

**Jan Larsen**
Technical University of Denmark, DTU Informatics, Bldg. 321, Richard Petersens Plads, DK-2800 Lyngby
jl@imm.dtu.dk

### Version History

| Version | Date | Change Revision |
|---|---|---|
| 1.0 | May 2010 | Initial issue. |

### Contents

## 1 Introduction

This paper acts as an appendix and supplement to the Journal paper entitled 'Unmixing of Hyperspectral Images using Bayesian Nonnegative Matrix Factorization with Volume Prior' [1]. It provides detailed derivation as a supplement to the Bayesian unmixing model.

If you should have any comments or corrections, please forward any inquiry directly to Morten Arngren (ma@imm.dtu.dk or info@arngren.com).

## 2 Bayesian NMF with volume prior

This section present the detailed derivations of the different probability distributions used in the Gibbs sampling scheme [1].

### 2.1 Posterior

The joint posterior distribution of the NMF parameters, $\boldsymbol{W}$ and $\boldsymbol{H}$, can be expressed using Bayes' rule

$$p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X},\mathcal{H}) = \frac{p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\sigma^2)p(\boldsymbol{H})p(\boldsymbol{W}|\gamma)p(\sigma^2|\alpha,\beta)}{p(\boldsymbol{X})}, \tag{1}$$

where $\mathcal{H}=\{\alpha,\beta,\gamma\}$ are hyperparameters and $\mathcal{P}=\{\boldsymbol{W},\boldsymbol{H},\sigma^2\}$ denote the parameters of the model.

### 2.2 Sampling the noise variance

The noise is assumed independent and identically distributed white Gaussian noise, giving rise to the following likelihood function,

$$p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\sigma^2) = \prod_{n=1}^{N}\prod_{m=1}^{M}\mathcal{N}\left(x_{mn}|\boldsymbol{w}_{m:}\boldsymbol{h}_{:n},\sigma^2\right) = \prod_{n=1}^{N}\prod_{m=1}^{M}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x_{mn}-\boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2}{-2\sigma^2}\right). \tag{2}$$

The noise variance $\sigma^2$ is modeled by the inverse gamma probability distribution given by

$$p(\sigma^2|\alpha,\beta) = \mathcal{IG}(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(-\frac{\beta}{\sigma^2}\right), \tag{3}$$

where $\alpha$ and $\beta$ are the parameters defining the distribution. The conditional probability of the noise variance can be expressed as

$$p(\sigma^2|\boldsymbol{X},\mathcal{P}_{\setminus\sigma^2},\mathcal{H}) = \frac{p(\sigma^2,\boldsymbol{X}|\mathcal{P}_{\setminus\sigma^2},\mathcal{H})}{p(\boldsymbol{X}|\mathcal{P}_{\setminus\sigma^2})} = \frac{p(\boldsymbol{X}|\mathcal{P}_{\setminus\sigma^2})p(\sigma^2|\mathcal{H})}{\int p(\sigma^2,\boldsymbol{X}|\mathcal{P}_{\setminus\sigma^2},\mathcal{H})\ \mathrm{d}\sigma^2} \tag{4}$$

$$\propto p(\boldsymbol{X}|\mathcal{P}_{\setminus\sigma^2})p(\sigma^2|\mathcal{H}) \tag{5}$$

$$= \prod_{n=1}^{N}\prod_{m=1}^{M}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x_{mn}-\boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2}{-2\sigma^2}\right)\right]\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(-\frac{\beta}{\sigma^2}\right). \tag{6}$$

The conjugate prior in Eq. (3) gives rise to a new set of parameters $\bar{\alpha}$ and $\bar{\beta}$ for the inverse gamma probability distribution. In estimating the new parameter set, constant terms are not considered leading to a reduction of Eq. (6).

$$p(\sigma^2|\boldsymbol{X},\mathcal{P}_{\setminus\sigma^2},\mathcal{H}) \propto \prod_{n=1}^{N}\prod_{m=1}^{M}\left[\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\right]\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(\sum_{n=1}^{N}\sum_{m=1}^{M}\frac{(x_{mn}-\boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2}{-2\sigma^2}\right)\exp\left(-\frac{\beta}{\sigma^2}\right) \tag{7}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}NM}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(\sum_{n=1}^{N}\sum_{m=1}^{M}\frac{(x_{mn}-\boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2}{-2\sigma^2}-\frac{\beta}{\sigma^2}\right) \tag{8}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1+\frac{1}{2}NM}\exp\left(-\frac{1}{\sigma^2}\left[\beta+\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}(x_{mn}-\boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2\right]\right) \tag{9}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\bar{\alpha}+1}\exp\left(-\frac{\bar{\beta}}{\sigma^2}\right). \tag{10}$$

The new set of hyperparameters are now given by

$$\bar{\alpha} = \alpha + \tfrac{1}{2}NM, \tag{11}$$

$$\bar{\beta} = \beta + \tfrac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}(x_{mn}-\boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2. \tag{12}$$

The inverse gamma distribution then becomes

$$p(\sigma^2|\boldsymbol{X},\mathcal{P}_{\setminus\sigma^2}) = \mathcal{IG}(\sigma^2|\bar{\alpha},\bar{\beta}), \tag{13}$$

### 2.3 Sampling the fractional abundances

The prior distribution of the fractional abundances is modeled as a uniform prior on the unit simplex expressed as

$$p(\boldsymbol{H}) \propto \prod_{n=1}^{N}\mathbb{I}\left[\sum_{k=1}^{K}h_{kn}=1\right]\prod_{k=1}^{K}\mathbb{I}\left[h_{kn}\geq 0\right], \tag{14}$$

$$p(\boldsymbol{h}_{:n}) \propto \mathbb{I}\left[\sum_{k=1}^{K} h_{kn} = 1\right] \prod_{k=1}^{K} \mathbb{I}\left[h_{kn} \geq 0\right], \tag{15}$$

The conditional probability distribution can then be calculated from the general expression as

$$\begin{aligned} p(\boldsymbol{h}_{:n}|\boldsymbol{X}, \mathcal{P}_{\backslash \boldsymbol{h}_{:n}}) &= \frac{p(\boldsymbol{W}, \boldsymbol{h}_{:n}|\boldsymbol{X}, \sigma^2)}{p(\boldsymbol{W}|\boldsymbol{X}, \sigma^2)} = \frac{p(\boldsymbol{W}, \boldsymbol{h}_{:n}|\boldsymbol{X}, \sigma^2)}{\int p(\boldsymbol{W}, \boldsymbol{h}_{:n}|\boldsymbol{X}, \sigma^2)\ \mathrm{d}\boldsymbol{h}_{:n}} && (16)\\ &= p(\boldsymbol{W}, \boldsymbol{h}_{:n}|\boldsymbol{X}, \sigma^2) \cdot d && (17)\\ &= \frac{p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{h}_{:n}, \sigma^2) p(\boldsymbol{h}_{:n}) p(\boldsymbol{W}|\gamma) p(\sigma^2|\alpha, \beta)}{p(\boldsymbol{X})} \cdot d && (18)\\ &\propto p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{h}_{:n}, \sigma^2) p(\boldsymbol{h}_{:n}) && (19)\\ &= \prod_{n=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(\boldsymbol{x}_{:n} - \boldsymbol{W}\boldsymbol{h}_{:n})^2}{-2\sigma^2}\right)\right] p(\boldsymbol{h}_{:n}), && (20) \end{aligned}$$

where $d$ denote any irrelevant constant terms. The argument of the exponential function can be rewritten in quadratic form if we denote the covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \mathcal{I}$, where $\mathcal{I}$ is the identity matrix.

$$\begin{aligned} &(\boldsymbol{x}_{:n} - \boldsymbol{W}\boldsymbol{h}_{:n})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_{:n} - \boldsymbol{W}\boldsymbol{h}_{:n}) = \\ &\qquad \boldsymbol{x}_{:n}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_{:n} - \boldsymbol{x}_{:n}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{h}_{:n} - \boldsymbol{h}_{:n}^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_{:n} + \boldsymbol{h}_{:n}^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{h}_{:n}. \end{aligned} \tag{21}$$

As Eq. (20) is also a distribution over the parameter $\boldsymbol{h}_{:n}$, it can be rewritten in quadratic form as

$$p(\boldsymbol{h}_{:n}|\boldsymbol{X}, \boldsymbol{W}, \sigma^2) = d \cdot \prod_{n=1}^{N} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left((\boldsymbol{h}_{:n} - \bar{\boldsymbol{\mu}}_n)^\top \bar{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{h}_{:n} - \bar{\boldsymbol{\mu}}_n)\right)\right] p(\boldsymbol{h}_{:n}), \tag{22}$$

where $d$ denote the new normalization constant. The mean offset is denoted by $\bar{\boldsymbol{\mu}}_n$ and $\bar{\boldsymbol{\Sigma}}$ is the covariance matrix for the fractional abundances $\boldsymbol{h}_{:n}$. These two new parameters are calculated by expanding the quadratic form in Eq. (22)

$$(\boldsymbol{h}_{:n} - \bar{\boldsymbol{\mu}}_n)^\top \bar{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{h}_{:n} - \bar{\boldsymbol{\mu}}_n) = \boldsymbol{h}_{:n}^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{h}_{:n} - \boldsymbol{h}_{:n}^\top \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_n - \bar{\boldsymbol{\mu}}_n^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{h}_{:n} + \bar{\boldsymbol{\mu}}_n^\top \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_n \tag{23}$$

Comparing Eq. (21) and (23), the new mean offset and covariance matrix can easily be identified as

$$\boldsymbol{h}_{:n}^\top \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{h}_{:n} = \boldsymbol{h}_{:n}^\top \underbrace{\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}}_{\bar{\boldsymbol{\Sigma}}^{-1}} \boldsymbol{h}_{:n} \qquad \Rightarrow \qquad \bar{\boldsymbol{\Sigma}} = (\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W})^{-1} = \sigma^2 (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \tag{24}$$

$$\begin{aligned} \boldsymbol{h}_{:n}^\top \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_n &= \boldsymbol{h}_{:n}^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_{:n} = \boldsymbol{h}_{:n}^\top \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\Sigma}} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_{:n} && (25)\\ &= \boldsymbol{h}_{:n}^\top \bar{\boldsymbol{\Sigma}}^{-1} \underbrace{(\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_{:n}}_{\bar{\boldsymbol{\mu}}_n} && (26)\\ \Rightarrow \qquad \bar{\boldsymbol{\mu}}_n &= (\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_{:n} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{x}_{:n}, && (27) \end{aligned}$$

where the last reduction step utilizes $\boldsymbol{\Sigma}^{-1}$ as a diagonal covariance matrix resulting in the pseudoinverse of $\boldsymbol{W}$. The new posterior conditional parameters are hence expressed by Eq. (24) and (27) and becomes

$$p(\boldsymbol{h}_{:n}|\boldsymbol{X},\mathcal{P}\backslash\boldsymbol{h}_{:n}) \propto \mathcal{N}(\boldsymbol{h}_{:n}|\bar{\boldsymbol{\mu}}_n,\bar{\boldsymbol{\Sigma}}_n)\prod_{n=1}^{N}\mathbb{I}\left[\sum_{k=1}^{K}h_{kn}=1\right]\prod_{k=1}^{K}\mathbb{I}\left[h_{kn}\geq 0\right], \tag{28}$$

### 2.4 Sampling endmembers

The prior probability of the endmembers are modeled as a constrained exponential distribution to encourage small spanning volumes of the endmembers and is expressed as

$$p(\boldsymbol{W}|\gamma) \propto \exp\big(-\gamma J_w(\boldsymbol{W})\big)\prod_{m=1}^{M}\prod_{k=1}^{K}\mathbb{I}\left[w_{mk}\geq 0\right], \tag{29}$$

where 3 different volume measures are considered given by

$$J_w^{\mathrm{pp}}(\boldsymbol{W}) = \det(\boldsymbol{W}^\top\boldsymbol{W}), \tag{30}$$

$$J_w^{\mathrm{sv}}(\boldsymbol{W}) = \det(\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}),\ \text{and} \tag{31}$$

$$J_w^{\mathrm{dist}}(\boldsymbol{W}) = \sum_{k=1}^{K}\left|\left|\boldsymbol{w}_{:k}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{:k'}\right|\right|_2^2. \tag{32}$$

The conditional prior distribution of a single endmember element then becomes a truncated Gaussian distribution due to the quadratic terms in any of the $J_w(\boldsymbol{W})$, i.e.

$$\begin{aligned} p(w_{mk}|\mathcal{P}_{\backslash w_{mk}}) &\propto \mathcal{N}(w_{mk}|c_{mk},s_{mk}^2)\mathbb{I}\left[w_{mk}\geq 0\right] \\ &= \frac{1}{\sqrt{2\pi}s_{mk}}\exp\left(\frac{(w_{mk}-c_{mk})^2}{-2s_{mk}^2}\right)\mathbb{I}\left[w_{mk}\geq 0\right], \end{aligned} \tag{33}$$

where $c_{mk}$ and $s_{mk}^2$ denote the corresponding set of parameters for the Gaussian distribution, mean and covariance respectively. Using the likelihood written in Eq. (2) the conditional posterior distribution then becomes

$$p(w_{mk}|\boldsymbol{X},\mathcal{P}_{\backslash w_{mk}}) = \frac{p(w_{mk},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{W}_{\backslash w_{mk}},\sigma^2)}{p(\boldsymbol{H}|\boldsymbol{X},\boldsymbol{W}_{\backslash w_{mk}}\sigma^2)} = \frac{p(w_{mk},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{W}_{\backslash w_{mk}}\sigma^2)}{\int p(w_{mk},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{W}_{\backslash w_{mk}}\sigma^2)\ \mathrm{d}w_{mk}} \tag{34}$$

$$= p(w_{mk},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{W}_{\backslash w_{mk}}\sigma^2)\cdot d \tag{35}$$

$$= \frac{p(\boldsymbol{X}|\mathcal{P})p(\boldsymbol{H})p(w_{mk}|\gamma)p(\sigma^2|\alpha,\beta)}{p(\boldsymbol{X})}\cdot d \tag{36}$$

$$\propto p(\boldsymbol{X}|\mathcal{P})p(w_{mk}|\gamma) \tag{37}$$

$$= \prod_{n=1}^{N}\prod_{m=1}^{M}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(x_{mn}-w_{mk}h_{kn})^2}{-2\sigma^2}\right)\right] \tag{38}$$

$$\times \exp\big(-\gamma J_w(w_{mk})\big)\mathbb{I}\left[w_{mk}\geq 0\right], \tag{39}$$

where $d$ denote any constant terms. As argued the conditional prior is a Gaussian distribution in Eq. (33) due to the quadratic terms and leads further to

$$p(w_{mk}|\boldsymbol{X}, \mathcal{P}_{\setminus w_{mk}}) \propto p(\boldsymbol{X}|\mathcal{P})p(w_{mk}|\mathcal{P}_{\setminus w_{mk}}) \tag{40}$$

$$= \prod_{n=1}^{N} \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left( \frac{(x_{mn} - \boldsymbol{w}_{m:}\boldsymbol{h}_{:n})^2}{-2\sigma^2} \right) \right] \frac{1}{\sqrt{2\pi}s_{mk}} \exp\left( \frac{(w_{mk} - c_{mk})^2}{-2s_{mk}^2} \right) \mathbb{I}\left[w_{mk} \geq 0\right] \tag{41}$$

$$= \frac{N}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}s_{mk}} \exp\left[ -\tfrac{1}{2}\sigma^{-2} \sum_{n=1}^{N} \left(x_{mn} - \boldsymbol{w}_{m:}\boldsymbol{h}_{:n}\right)^2 - \tfrac{1}{2}s_{mk}^{-2}\left(w_{mk} - c_{mk}\right)^2 \right] \mathbb{I}\left[w_{mk} \geq 0\right] \tag{42}$$

Using the argument of the exponential function in quadratic form, we get

$$-\tfrac{1}{2}\sigma^{-2} \sum_{n=1}^{N} \left(x_{mn} - \boldsymbol{w}_{m:}\boldsymbol{h}_{:n}\right)^2 - \tfrac{1}{2}s_{mk}^{-2}\left(w_{mk} - c_{mk}\right)^2 \tag{43}$$

$$= -\tfrac{1}{2}\sigma^{-2}\left(\boldsymbol{x}_{m:} - \boldsymbol{w}_{m:}\boldsymbol{H}\right)\left(\boldsymbol{x}_{m:} - \boldsymbol{w}_{m:}\boldsymbol{H}\right)^\top - \tfrac{1}{2}s_{mk}^{-2}\left(w_{mk} - c_{mk}\right)^2 \tag{44}$$

$$= -\tfrac{1}{2}\sigma^{-2}\left(\boldsymbol{x}_{m:} - w_{mk}\boldsymbol{h}_{k:} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:}\right)\left(\boldsymbol{x}_{m:} - w_{mk}\boldsymbol{h}_{k:} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:}\right)^\top - \tfrac{1}{2}s_{mk}^{-2}\left(w_{mk} - c_{mk}\right)^2 \tag{45}$$

$$= -\tfrac{1}{2}\sigma^{-2}\left(w_{mk}^2\boldsymbol{h}_{k:}\boldsymbol{h}_{k:}^\top - 2w_{mk}\boldsymbol{h}_{k:}(\boldsymbol{x}_{m:} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:})^\top + d\right) - \tfrac{1}{2}s_{mk}^{-2}\left(w_{mk} - c_{mk}\right)^2 \tag{46}$$

$$\propto\ -\sigma^{-2}\boldsymbol{h}_{k:}\boldsymbol{h}_{k:}^\top w_{mk}^2 + 2\sigma^{-2}w_{mk}\boldsymbol{h}_{k:}(\boldsymbol{x}_{m:} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:})^\top - s_{mk}^{-2}w_{mk}^2 + 2s_{mk}^{-2}c_{mk}w_{mk}, \tag{47}$$

where $d$ represent the constant terms. The conditional posterior is also a Gaussian distribution and can hence be expressed in quadratic form similarly to Eq. (23)

$$-\bar{\sigma}_{mk}^{-2}(w_{mk} - \bar{\mu}_{mk})^2 = -\bar{\sigma}_{mk}^{-2}w_{mk}^2 + 2\bar{\sigma}_{mk}^{-2}w_{mk}\bar{\mu}_{mk} - \bar{\sigma}_{mk}^{-2}\bar{\mu}_{mk}^2. \tag{48}$$

Both parameters can now be derived

$$-\bar{\sigma}_{mk}^{-2}w_{mk}^2 = -\sigma^{-2}\boldsymbol{h}_{k:}\boldsymbol{h}_{k:}^\top w_{mk}^2 - s_{mk}^{-2}w_{mk}^2 \qquad \Leftrightarrow \tag{49}$$

$$\bar{\sigma}_{mk}^{-2} = s_{mk}^{-2} + \sigma^{-2}\boldsymbol{h}_{k:}\boldsymbol{h}_{k:}^\top. \tag{50}$$

$$2\bar{\sigma}_{mk}^{-2}w_{mk}\bar{\mu}_{mk} = 2\sigma^{-2}w_{mk}\boldsymbol{h}_{k:}(\boldsymbol{x}_{m:} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:})^\top + 2s_{mk}^{-2}c_{mk}w_{mk} \qquad \Leftrightarrow \tag{51}$$

$$\bar{\sigma}_{mk}^{-2}\bar{\mu}_{mk} = \sigma^{-2}\boldsymbol{h}_{k:}(\boldsymbol{x}_{m:} - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:})^\top + s_{mk}^{-2}c_{mk} \qquad \Leftrightarrow \tag{52}$$

$$\bar{\mu}_{mk} = \bar{\sigma}_{mk}^2\Big(c_{mk}s_{mk}^{-2} + (\boldsymbol{x}_{m:}\boldsymbol{h}_{k:}^\top - \boldsymbol{w}_{m\bar{k}}\boldsymbol{h}_{\bar{k}:}\boldsymbol{h}_{k:}^\top)\sigma^{-2}\Big). \tag{53}$$

The parameters of the conditional Gaussian distribution is hence given by Eq. (50) and (53). The mean $c_{mk}$ and covariance $s_{mk}^{-2}$ of the conditional prior can now be determined for each of the volume measures in Eq. (30)-(32). These parameter are calculated from the conditional prior distribution rewritten as

$$p(w_{mk}|\mathcal{P}_{\setminus w_{mk}}) \propto \frac{1}{\sqrt{2\pi}s_{mk}} \exp\Big(-\gamma J_w(\boldsymbol{W})\Big)\mathbb{I}\left[w_{mk} \geq 0\right]. \tag{54}$$

8

In order to find the equivalent mean and covariance for the determinant based volume measures the argument of the determinant must be decomposed into scalar values using the following expression for the determinant of a symmetric matrix,

$$\det \begin{bmatrix} a & \boldsymbol{b}^\top \\ \boldsymbol{b} & \boldsymbol{C} \end{bmatrix} = a \det \left( \boldsymbol{C} - \tfrac{1}{a} \boldsymbol{b}\boldsymbol{b}^\top \right) \tag{55}$$

$$= a \det(\boldsymbol{C}) \left(1 - \tfrac{1}{a} \boldsymbol{b}^\top \boldsymbol{C}^{-1} \boldsymbol{b}\right) \tag{56}$$

$$= a \det(\boldsymbol{C}) - \boldsymbol{b}^\top \mathrm{adj}(\boldsymbol{C}) \boldsymbol{b}, \tag{57}$$

where Eq. (55) is the expression for the determinant of a block matrix, Eq. (56) follows from the matrix determinant lemma, and in Eq. (57) we have used the definition of the matrix adjugate.

### *2.4.1 Parellelepiped Volume Measure*

The parallelepipped volume given in Eq. (30) is decomposed using Eq. (57) to isolate $w_{mk}$ for the conditional prior distribution.

$$\det(\boldsymbol{W}^\top \boldsymbol{W}) = \det \left( \begin{bmatrix} w_{mk} & \boldsymbol{w}_{m\tilde{k}} \\ \boldsymbol{w}_{\tilde{m}k} & \boldsymbol{W}_{\tilde{m}\tilde{k}} \end{bmatrix}^\top \begin{bmatrix} w_{mk} & \boldsymbol{w}_{m\tilde{k}} \\ \boldsymbol{w}_{\tilde{m}k} & \boldsymbol{W}_{\tilde{m}\tilde{k}} \end{bmatrix} \right) \tag{58}$$

$$= \det \left( \begin{bmatrix} w_{mk} & \boldsymbol{w}_{\tilde{m}k}^\top \\ \boldsymbol{w}_{m\tilde{k}}^\top & \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \end{bmatrix} \begin{bmatrix} w_{mk} & \boldsymbol{w}_{m\tilde{k}} \\ \boldsymbol{w}_{\tilde{m}k} & \boldsymbol{W}_{\tilde{m}\tilde{k}} \end{bmatrix} \right) \tag{59}$$

$$= \det \left( \begin{bmatrix} w_{mk}^2 + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{w}_{\tilde{m}k} & w_{mk}\boldsymbol{w}_{m\tilde{k}} + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{W}_{\tilde{m}\tilde{k}} \\ \boldsymbol{w}_{m\tilde{k}}^\top w_{mk} + \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} & \boldsymbol{w}_{m\tilde{k}}^\top \boldsymbol{w}_{m\tilde{k}} + \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{W}_{\tilde{m}\tilde{k}} \end{bmatrix} \right) \tag{60}$$

$$= \det \left( \begin{bmatrix} w_{mk}^2 + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{w}_{\tilde{m}k} & w_{mk}\boldsymbol{w}_{m\tilde{k}} + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{W}_{\tilde{m}\tilde{k}} \\ \boldsymbol{w}_{m\tilde{k}}^\top w_{mk} + \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} & \boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}} \end{bmatrix} \right) \tag{61}$$

This leads to the specific definition of the parameters $a$, $\boldsymbol{b}$ and $\boldsymbol{C}$ to be

$$a = w_{mk}^2 + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{w}_{\tilde{m}k} \tag{62}$$

$$\boldsymbol{b} = \boldsymbol{w}_{m\tilde{k}}^\top w_{mk} + \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} \tag{63}$$

$$\boldsymbol{C} = \boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}. \tag{64}$$

Using Eq. (57) we get

$$\begin{aligned}
\gamma \det(\boldsymbol{W}^\top \boldsymbol{W}) &= \gamma \left( w_{mk}^2 + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{w}_{\tilde{m}k} \right) \det(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \\
&\quad - \gamma \left( w_{mk} \boldsymbol{w}_{m\tilde{k}} + \boldsymbol{w}_{\tilde{m}k}^\top \boldsymbol{W}_{\tilde{m}\tilde{k}} \right) \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \left( \boldsymbol{w}_{m\tilde{k}}^\top w_{mk} + \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} \right) && (65) \\
&= w_{mk}^2 \gamma \det(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) - \gamma w_{mk} \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{w}_{m\tilde{k}}^\top w_{mk} \\
&\quad - 2\gamma w_{mk} \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} + d && (66) \\
&= w_{mk}^2 \left( \gamma \det(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) - \gamma \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{w}_{m\tilde{k}}^\top \right) \\
&\quad - 2\gamma w_{mk} \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} + d, && (67)
\end{aligned}$$

where $d$ denote the irrelevant constant terms not including the variable $w_{mk}$. The quadratic form of the argument of the exponential function in Eq. (33) is expanded to find the new parameters

$$s_{mk}^{-2}(w_{mk} - c_{mk})^2 = s_{mk}^{-2} w_{mk}^2 - 2 s_{mk}^{-2} w_{mk} c_{mk} + s_{mk}^{-2} c_{mk}^2 \tag{68}$$

The new mean and covariance can now easily be identified

$$\begin{aligned}
s_{mk}^{-2} w_{mk}^2 &= w_{mk}^2 \left( \gamma \det(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) - \gamma \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{w}_{m\tilde{k}}^\top \right) \quad \Leftrightarrow && (69) \\
s_{mk}^{-2} &= \gamma \left( \det(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) - \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{w}_{m\tilde{k}}^\top \right). && (70)
\end{aligned}$$

$$\begin{aligned}
-2 s_{mk}^{-2} w_{mk} c_{mk} &= -2\gamma w_{mk} \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} \quad \Leftrightarrow && (71) \\
c_{mk} &= \gamma s_{mk}^2 \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k}. && (72)
\end{aligned}$$

*2.4.2 Simplex Volume Measure*

The simplex volume expressed in Eq. (31) is also decomposed using Eq. (57) to isolate $w_{mk}$ for the conditional prior distribution.

$$\begin{aligned}
\det(\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}) &= \det \left( \begin{bmatrix} w_{mk} - w_{m\rho} & \boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1} \\ \boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho} & \boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1} \end{bmatrix}^\top \begin{bmatrix} w_{mk} - w_{m\rho} & \boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1} \\ \boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho} & \boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1} \end{bmatrix} \right) && (73) \\
&= \det \left( \begin{bmatrix} w_{mk} - w_{m\rho} & (\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho})^\top \\ (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1})^\top & (\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1})^\top \end{bmatrix} \begin{bmatrix} w_{mk} - w_{m\rho} & \boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1} \\ \boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho} & \boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1} \end{bmatrix} \right) && (74) \\
&= \det \left( \begin{bmatrix} (w_{mk} - w_{m\rho})^2 & (w_{mk} - w_{m\rho})(\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1}) \\ +(\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho})^\top (\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho}) & +(\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho})^\top (\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1}) \\ (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1})^\top (w_{mk} - w_{m\rho}) & (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1})^\top (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1}) \\ +(\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1})^\top (\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho}) & +(\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1})^\top (\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1}) \end{bmatrix} \right). && (75)
\end{aligned}$$

This leads to the specific definition of the parameters $a$, $\boldsymbol{b}$ and $\boldsymbol{C}$ to be

$$a = (w_{mk} - w_{m\rho})^2 + (\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho})^\top (\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho}) \tag{76}$$

$$= (w_{mk} - w_{m\rho})^2 + \tilde{\boldsymbol{w}}_{\tilde{m}k}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k} \tag{77}$$

$$\boldsymbol{b} = (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1})^\top (w_{mk} - w_{m\rho}) + (\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1})^\top (\boldsymbol{w}_{\tilde{m}k} - \boldsymbol{w}_{\tilde{m}\rho}) \tag{78}$$

$$= \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top (w_{mk} - w_{m\rho}) + \tilde{\boldsymbol{W}}_{\tilde{m}\tilde{k}}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k} \tag{79}$$

$$\boldsymbol{b}^\top = (w_{mk} - w_{m\rho})\tilde{\boldsymbol{w}}_{m\tilde{k}} + \tilde{\boldsymbol{w}}_{\tilde{m}k}^\top \tilde{\boldsymbol{W}}_{\tilde{m}\tilde{k}} \tag{80}$$

$$\boldsymbol{C} = (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1})^\top (\boldsymbol{w}_{m\tilde{k}} - w_{m\rho}\mathbf{1}) + (\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1})^\top (\boldsymbol{W}_{\tilde{m}\tilde{k}} - \boldsymbol{w}_{\tilde{m}\rho}\mathbf{1}) \tag{81}$$

$$= (\boldsymbol{W}_{m\tilde{k}} - \boldsymbol{w}_{:\rho}\mathbf{1})^\top (\boldsymbol{W}_{m\tilde{k}} - \boldsymbol{w}_{:\rho}\mathbf{1}) \tag{82}$$

$$= \tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}} \tag{83}$$

Applying Eq. (57) we get

$$\gamma \det(\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}) = \gamma \left( (w_{mk} - w_{m\rho})^2 + \tilde{\boldsymbol{w}}_{\tilde{m}k}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k} \right) \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}})$$
$$- \gamma \left( (w_{mk} - w_{m\rho})\tilde{\boldsymbol{w}}_{m\tilde{k}} + \tilde{\boldsymbol{w}}_{\tilde{m}k}^\top \tilde{\boldsymbol{W}}_{\tilde{m}\tilde{k}} \right) \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \left( \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top (w_{mk} - w_{m\rho}) + \tilde{\boldsymbol{W}}_{\tilde{m}\tilde{k}}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k} \right) \tag{84}$$

$$= (w_{mk} - w_{m\rho})^2 \gamma \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) - \gamma (w_{mk} - w_{m\rho}) \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top (w_{mk} - w_{m\rho})$$
$$- 2\gamma (w_{mk} - w_{m\rho}) \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k} + d_1 \tag{85}$$

$$= w_{mk}^2 \gamma \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) - 2 w_{mk} w_{m\rho} \gamma \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}})$$
$$- \gamma w_{mk}^2 \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top + 2\gamma w_{mk} w_{m\rho} \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top$$
$$- 2\gamma w_{mk} \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \tilde{\boldsymbol{w}}_{\tilde{m}k} + d_2 \tag{86}$$

$$= w_{mk}^2 \left( \gamma \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) - \gamma \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top \right)$$
$$- w_{mk} \Big( 2 w_{m\rho} \gamma \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) - 2\gamma w_{m\rho} \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top$$
$$+ 2\gamma \boldsymbol{w}_{m\tilde{k}} \mathrm{adj}(\boldsymbol{W}_{m\tilde{k}}^\top \boldsymbol{W}_{m\tilde{k}}) \boldsymbol{W}_{\tilde{m}\tilde{k}}^\top \boldsymbol{w}_{\tilde{m}k} \Big) + d_2, \tag{87}$$

where $\{d_i\}_{i=1}^2$ denote the irrelevant constant terms not including the variable $w_{mk}$. The new mean and covariance can now easily be derived using the quadratic form in Eq. (68).

$$s_{mk}^{-2} w_{mk}^2 = w_{mk}^2 \left( \gamma \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) - \gamma \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top \right) \quad \Leftrightarrow$$
$$s_{mk}^{-2} = \gamma \left( \det(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) - \tilde{\boldsymbol{w}}_{m\tilde{k}} \mathrm{adj}(\tilde{\boldsymbol{W}}_{m\tilde{k}}^\top \tilde{\boldsymbol{W}}_{m\tilde{k}}) \tilde{\boldsymbol{w}}_{m\tilde{k}}^\top \right), \tag{88}$$

$$-2s_{mk}^{-2}w_{mk}c_{mk} = -w_{mk}\Big(2w_{m\rho}\gamma\det(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}}) - 2\gamma w_{m\rho}\tilde{\boldsymbol{w}}_{m\bar{k}}\mathrm{adj}(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}})\tilde{\boldsymbol{w}}_{m\bar{k}}^{\top} + 2\gamma\boldsymbol{w}_{m\bar{k}}\mathrm{adj}(\boldsymbol{W}_{m\bar{k}}^{\top}\boldsymbol{W}_{m\bar{k}})\boldsymbol{W}_{\tilde{m}\bar{k}}^{\top}\boldsymbol{w}_{\tilde{m}k}\Big) \quad \Leftrightarrow \tag{89}$$

$$c_{mk} = s_{mk}^{2}\Big(w_{m\rho}\gamma\det(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}}) - \gamma w_{m\rho}\tilde{\boldsymbol{w}}_{m\bar{k}}\mathrm{adj}(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}})\tilde{\boldsymbol{w}}_{m\bar{k}}^{\top} + \gamma\boldsymbol{w}_{m\bar{k}}\mathrm{adj}(\boldsymbol{W}_{m\bar{k}}^{\top}\boldsymbol{W}_{m\bar{k}})\boldsymbol{W}_{\tilde{m}\bar{k}}^{\top}\boldsymbol{w}_{\tilde{m}k}\Big) \quad \Leftrightarrow \tag{90}$$

$$c_{mk} = \frac{w_{m\rho}\gamma\det(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}}) - \gamma w_{m\rho}\tilde{\boldsymbol{w}}_{m\bar{k}}\mathrm{adj}(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}})\tilde{\boldsymbol{w}}_{m\bar{k}}^{\top}}{\gamma\det(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}}) - \gamma\tilde{\boldsymbol{w}}_{m\bar{k}}\mathrm{adj}(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}})\tilde{\boldsymbol{w}}_{m\bar{k}}^{\top}} \cdot s_{mk}^{2}\gamma\boldsymbol{w}_{m\bar{k}}\mathrm{adj}(\boldsymbol{W}_{m\bar{k}}^{\top}\boldsymbol{W}_{m\bar{k}})\boldsymbol{W}_{\tilde{m}\bar{k}}^{\top}\boldsymbol{w}_{\tilde{m}k} \quad \Leftrightarrow \tag{91}$$

$$c_{mk} = \frac{w_{m\rho}\gamma\det(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}}) - \gamma w_{m\rho}\tilde{\boldsymbol{w}}_{m\bar{k}}\mathrm{adj}(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}})\tilde{\boldsymbol{w}}_{m\bar{k}}^{\top}}{\gamma\det(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}}) - \gamma\tilde{\boldsymbol{w}}_{m\bar{k}}\mathrm{adj}(\tilde{\boldsymbol{W}}_{m\bar{k}}^{\top}\tilde{\boldsymbol{W}}_{m\bar{k}})\tilde{\boldsymbol{w}}_{m\bar{k}}^{\top}} \cdot s_{mk}^{2}\gamma\boldsymbol{w}_{m\bar{k}}\mathrm{adj}(\boldsymbol{W}_{m\bar{k}}^{\top}\boldsymbol{W}_{m\bar{k}})\boldsymbol{W}_{\tilde{m}\bar{k}}^{\top}\boldsymbol{w}_{\tilde{m}k} \quad \Leftrightarrow \tag{92}$$

$$c_{mk} = w_{m\rho} + s_{mk}^{2}\gamma\boldsymbol{w}_{m\bar{k}}\mathrm{adj}(\boldsymbol{W}_{m\bar{k}}^{\top}\boldsymbol{W}_{m\bar{k}})\boldsymbol{W}_{\tilde{m}\bar{k}}^{\top}\boldsymbol{w}_{\tilde{m}k} \tag{93}$$

The derived mean and covariance shows how using the relative simplex vectors $\tilde{\boldsymbol{W}}$ leads to the intuitive result of only adding the offset $m_{m\rho}$ to the mean $c_{mk}$.

### *2.4.3 Euclidian Distance Measure*

The volume measure based on euclidian distance metric can be derived by isolating $w_{mk}$ in Eq. (32).

$$\gamma\sum_{k=1}^{K}\left|\left|\boldsymbol{w}_{:k}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{:k'}\right|\right|_2^2 = \sum_{\bar{k}=1}^{K}\left[\left(w_{m\bar{k}}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{mk'}\right)^2 + \sum_{\tilde{m}\neq m}\left(\boldsymbol{w}_{\tilde{m}\bar{k}}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{\tilde{m}k'}\right)^2\right] \tag{94}$$

$$= \sum_{\bar{k}=1}^{K}\left(w_{m\bar{k}}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{mk'}\right)^2 + d_1 \tag{95}$$

$$= \left(w_{mk}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{mk'}\right)^2 + \sum_{\bar{k}\neq k}\left(w_{m\bar{k}}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{mk'}\right)^2 + d_1 \tag{96}$$

$$= \left(w_{mk}-\frac{1}{K}\sum_{k'=1}^{K}\boldsymbol{w}_{mk'}\right)^2 + d_2 \tag{97}$$

$$= \left(w_{mk}-\frac{1}{K}w_{mk}-\frac{1}{K}\sum_{k'\neq k}\boldsymbol{w}_{mk'}\right)^2 + d_2 \tag{98}$$

$$= \left(\frac{K-1}{K}w_{mk}-\frac{1}{K}\sum_{k'\neq k}\boldsymbol{w}_{mk'}\right)^2 + d_2 \tag{99}$$

$$= \left(\frac{K-1}{K}\right)^2\left(w_{mk}-\frac{1}{K-1}\sum_{k'\neq k}\boldsymbol{w}_{mk'}\right)^2 + d_2 \tag{100}$$

where $\{d_i\}_{i=1}^2$ denote the irrelevant constant terms not including the variable $w_{mk}$. The mean and covariance can now easily be derived as

$$s_{mk}^{-2} = \gamma \left( \frac{K-1}{K} \right)^2, \tag{101}$$

$$c_{mk} = \frac{1}{K-1} \sum_{k' \neq k} w_{mk'}. \tag{102}$$

## 3 Comments

If you should have any comments or corrections, please forward any inquiry directly to Morten Arngren (ma@imm.dtu.dk or info@arngren.com).

## References

1. Arngren, M., Schmidt, M.N., Larsen, J.: Unmixing of hyperspectral images using bayesian nonnegative matrix factorization with volume prior. Journal of Signal Processing Systems (2010)

Appendix E

# Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis

Authors: Morten Arngren, Jan Larsen, Per Waaben Hansen, Birger Eriksen, and Rasmus Larsen.

Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis. *Journal of Agricultural and Food Chemistry, vol. 59 (21), p. 11385-11394, DOI: 10.1021/jf202122y, 2011.* Published.

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

ARTICLE

pubs.acs.org/JAFC

# Analysis of Pregerminated Barley Using Hyperspectral Image Analysis

Morten Arngren,[*,†,‡] Per Waaben Hansen,[*,‡] Birger Eriksen,[*,§] Jan Larsen,[*,†] and Rasmus Larsen[*,†]

[†]DTU Informatics, Technical University of Denmark, Richard Petersens Plads Bldg 321, DK-2800 Kgs. Lyngby, Denmark

[‡]FOSS Analytical A/S, Slangerupgade 69, DK-3400 Hillerød, Denmark

[§]Sejet Planteforædling I/S, Nørremarksvej 67, DK-8700 Horsens, Denmark

**ABSTRACT:** Pregermination is one of many serious degradations to barley when used for malting. A pregerminated barley kernel can under certain conditions not regerminate and is reduced to animal feed of lower quality. Identifying pregermination at an early stage is therefore essential in order to segregate the barley kernels into low or high quality. Current standard methods to quantify pregerminated barley include visual approaches, e.g. to identify the root sprout, or using an embryo staining method, which use a time-consuming procedure. We present an approach using a near-infrared (NIR) hyperspectral imaging system in a mathematical modeling framework to identify pregerminated barley at an early stage of approximately 12 h of pregermination. Our model only assigns pregermination as the cause for a single kernel's lack of germination and is unable to identify dormancy, kernel damage etc. The analysis is based on more than 750 Rosalina barley kernels being pregerminated at 8 different durations between 0 and 60 h based on the BRF method. Regerminating the kernels reveals a grouping of the pregerminated kernels into three categories: normal, delayed and limited germination. Our model employs a supervised classification framework based on a set of extracted features insensitive to the kernel orientation. An out-of-sample classification error of 32% ($CI_{95\%}$: 29−35%) is obtained for single kernels when grouped into the three categories, and an error of 3% ($CI_{95\%}$: 0−15%) is achieved on a bulk kernel level. The model provides class probabilities for each kernel, which can assist in achieving homogeneous germination profiles. This research can further be developed to establish an automated and faster procedure as an alternative to the standard procedures for pregerminated barley.

**KEYWORDS:** barley, pregermination, BRF method, near-infrared, NIR, hyperspectral camera, multinomial classification

## INTRODUCTION

Barley is a cereal grain and is the cereal crop ranking number four in the world both in terms of quantity produced and in area of cultivation. Barley has many uses. It serves as a major animal fodder, as base malt for beer and certain distilled beverages, and as an ingredient in various food products. Only high quality barley is used for malting and beer production. Grain lots not fulfilling the quality requirements for malting will be downgraded to feed barley, which has lower economic value. The quality is defined by many different parameters, e.g. protein level, grading, dormancy, pregermination, physical damage in terms of inability to germinate, or even identification of various fungal infections. Identifying pregerminated barley in particular has so far been one of the most difficult quality parameters to handle both for the malting/brewing industry and for farmers/grain companies.

Pregermination (preharvest sprouting) is a major problem in seasons with wet harvest conditions. Germination will be initiated in the field and later stop again because of a change in weather conditions or drying of the grain after harvest and before storing. The barley may have little or no visible indication of the pregermination. The major defect of pregerminated barley is its predisposed inability to germinate again, e.g. for malting. When the germination reaches a certain level and is afterward stopped, the kernel can simply not continue the germination or will germinate with a reduced speed and vigor. A different limitation is kernels having dormancy as they cannot germinate due to a chemical lock preventing the germination process to start. This trait is hence sometimes considered as a positive trait as is hinders pregermination, but can require additional storing time to overcome the dormancy prior to malting.

Currently the germination ability of barley kernels can be evaluated using different approaches. The kernels can be subjected to a simple germination process run over a few days and with relatively few kernels by applying for instance the BRF method.[1] Specifically to determine pregermination, different staining methods can be used.[2,3] Visual inspection is also widely used, but it is subjective and lacks the sensitivity to detect sprouting early.[4] Common to these approaches is their high time consumption. Any method able to identify pregerminated barley within minutes or even classify the individual kernels into grades of quality will provide a tool for additional quality assessment to the industry. Basing such a system on computer vision provides the fast and nondestructive approach called for.

The research in pregerminated cereals has received attention in research communities to address the issue of identifying pregerminated barley kernels using automated visual systems. The simplest systems use cameras operating in the visual spectral range providing a high resolution. These image systems commonly rely on the spatial information and are often only able to detect germination by identifying a sprout on the kernels or to detect kernel deformations.[5] Takeuchi et al.[6] used such a computer vision system to detect morphological changes of barley kernels during germination based on 50 barley kernels.

The visual spectral range lies between 400 and 800 nanometers (nm) and has the limitation of only acting on the surface of

Received: May 31, 2011
Accepted: September 21, 2011
Revised: September 21, 2011
Published: September 21, 2011

Journal of Agricultural and Food Chemistry ARTICLE

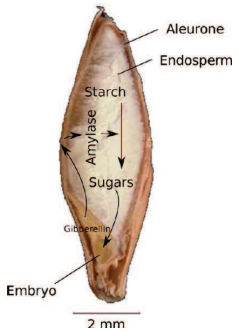



**Figure 1.** The internal biochemical mechanisms of a barley kernel during germination (courtesy of KAMPFFMEYER Food Innovation GmbH).

the kernels with insignificant penetration through the husk. Thus only identification of pregermination from external features is possible. As the initial process of germination is occurring inside at the embryo and endosperm, such an approach potentially leads to delayed detection. Spatial features used in visual systems can further be sensitive toward the orientation of the kernel.

A natural extension is to include near-infrared (NIR) wavelengths from about 800 to 2500 nm, where the light has the capability to penetrate the kernels through the husk and extract internal image information. NIR imaging thus has the potential to extract chemical changes before they cause an effect on the outside of the kernel. This approach can further lead to early detection of germination and can be superior to visual based systems. Xing et al.[5] combined the visual and NIR spectral ranges (400−1000 nm) to detect sprout damage to Canada Western Red Spring wheat. They conducted classification on approximately 200 wheat kernels germinated for 0, 24 and 48 h by exploiting various spatial changes of the kernels during germination.

Many applications using the visual and NIR spectral range are based on fewer than 20 different wavelengths. This spectral resolution may not be sufficient in applications, where similar constituents need to be separated. Hyperspectral imaging provides a high spectral resolution with more than 100 bands, usually at the expense of a lower spatial resolution. This allows for models to utilize the spectral information, possibly in combination with the image data. Xing et al.[7] exploited this combination to classify $\alpha$-amylase activity in individual Canadian western wheat kernels into two classes using hyperspectral NIR images. The predictions were based on the partial least-squares (PLS) model using mean spectra from the image segmented embryo leading to classification rates above 80%.

Munck and Møller[8] had developed a model based on NIR spectroscopy with separately acquired image information to develop a model able to predict the 24 h germination of barley kernels. Engelbrecht et al.[9] illustrated the potential of using hyperspectral imaging from 1000 to 2500 nm to classify barley into normal or germinated kernels based on 150 grains germinated for 6, 9, 12, 18 and 24 h. Their work further included the destructive tetrazolium test to obtain single kernel reference values on their viability. Manley et al.[10] extended the work to

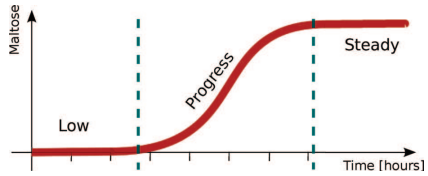



**Figure 2.** Conceptual illustration of the progression of germinating barley showing the starch breakdown. The two thresholds indicate the start and stop of the major chemical changes inside the kernel during germination.

employ a PLS discriminant analysis model to classify single kernels into normal or pregerminated and achieve a classification error of 2.87%.

Analyzing barley using NIR imaging by building a mathematical model to provide the degree of pregermination of single kernels with true pregermination times as references has not been attempted previously. We therefore investigate if a mathematical model can be developed for the detection of pregerminated barley and for describing the degree of pregermination of a single kernel using hyperspectral NIR imaging with true pregermination time as reference.

In the next section we give an introduction to barley germination in general, including the pregermination experiment followed by a description of the camera system, the preprocessing pipeline and the theoretical background for our analysis and model choice. The classification results are presented and discussed in the Results. A summary with discussions is presented in the last section. Supplementary details concerning the image acquisition and the models used are described in a supplementary technical report.[11]

## MATERIALS AND METHODS

**Barley Germination.** A barley kernel (*Hordeum vulgare*) is a member of the grass family and has internal structure as illustrated in Figure 1. Germination of barley is the process of the grain starting to sprout and becoming a plant. Biologically the process is highly complex with many stages, and in this context only the general steps will be considered. It partly involves the conversion of starch into sugars (maltose and glucose) used for the growth. The germination process is initiated by the absorption of water into the kernel. Provided the environmental conditions are suitable, e.g. temperature, this starts a whole series of biological steps including the release of the enzyme $\alpha$-amylase from the aleurone layer into the endosperm, as shown in Figure 1. The enzyme breaks the starch molecules down into oliosaccharides, subsequently to the disaccharide maltose, and finally to glucose, which is transported to the embryo for radicle and epicotyl growth. This activity can be seen as a large scale physical change inside the kernel. A conceptual illustration of the progression of the germination in barley is depicted in Figure 2. It starts with a short period with little activity, where the $\alpha$-amylase release is commencing, then rises as the amylase activity is maximum and finally settles as all the starch is completely broken down.

**Pregermination Experiment.** An experiment was set up in collaboration with Sejet Planteforaedling, Denmark, to produce barley kernels of the Rosalina variety with eight different durations of pregermination from 0 to 60 h. The experiment was conducted using the following procedure.

Journal of Agricultural and Food Chemistry ARTICLE

**Table 1. Number of Kernels Used in the Different Steps in the Analysis and the Concentration of Water in the Kernels with an Accuracy of 0.25%age Points**[a]

| | 0 h[b] | 12 h[b] | 18 h[b] | 24 h[b] | 30 h[b] | 36 h[b] | 48 h[b] | 60 h[b] | total |
|---|---|---|---|---|---|---|---|---|---|
| pregerminated | 2000 | 500 | 500 | 2000 | 2000 | 2000 | 500 | 500 | 10000 |
| regerminated | 300 | 300 | 300 | 1000 | 300 | 300 | 300 | 300 | 3100 |
| image acquired | 150 | 97 | 50 | 150 | 150 | 150 | 47 | 46 | 840 |
| preprocessed | 135 | 87 | 45 | 135 | 135 | 135 | 42 | 41 | 755 |
| water | 10.4% | 10.2% | 9.9% | 10.2% | 10.2% | 10.1% | 10.0% | 9.8% | |

[a] Since not all germination times had the same number of kernels, only images of a subset of the kernels were acquired. [b] Germination time.

1. Two filter papers (upper and lower side) were soaked in water for a few hours and afterward left to drip off. Prior to use, the filter paper must still be wet, but not soaking or dripping.
2. 100 single kernels were placed on the lower filter paper.
3. The upper filter paper was placed on top and the enclosed kernels were rolled up to a small cylinder locked with a rubber band and placed in a dark and warm environment (e.g., a cardboard box at room temperature).
4. When the germination times were reached, the individual kernels were immediately put in a net bag and placed in front of a blow dryer for a minimum of three hours at 5 to 10 °C.

500 kernels were pregerminated for each of the times [12, 18, 48, 60 h], and 2000 kernels were pregerminated for each of the times [0, 24, 30, 36 h] in order to conduct a reliable statistical analysis in the time range, where the kernels would start and stop the major chemical change (cf. Figure 2). When the moisture level inside the kernels drops below approximately 35% during the drying process, the germination process is considered halted. The kernels were further dried to a water content of approximately 7−27%, which is considered as a stable level for long-term storage.

After the kernels had pregerminated and had been dried by blow drying, they were stored in a refrigerator for approximately 6 months. Over 3000 of the kernels were then subjected to a regermination experiment (described below) to measure the degrading effect of the pregermination. Including wastage this left 6200 kernels for an additional 4 months of storage before the hyperspectral images were taken. Since not all germination times had the same number of kernels, only images of a subset of the kernels were acquired, i.e. 840 kernels (cf. Table 1). This long-term storage slowly dried the kernels further and led to equal water contents of approximately 10% between the pregermination times using FOSS NIR XDS instrument analysis (cf. Table 1).

Prior to the hyperspectral image acquisition, the epicotyl and radicle were rubbed off to fit the kernels on the sample plate. This has no impact on our analysis as we attempt to capture the pregermination inside the kernels using penetrating light (NIR) instead of relying on visible physical changes. During the image acquisition, the germination of every kernel was verified by visual inspection. This is, however, associated with some uncertainty as germination may not be visible for the short germination times. The germination time references designate the period the kernels were inside the germination environment, i.e. from when the filter paper is locked with rubber bands to the start of the dry blowing process. The periods obviously become a bit longer, as the germination will only halt when the water content falls below approximately 35%. This additional germination time is considered equal for all periods and is hence not included in the germination time references.

**Germinative Energy of Barley.** The degrading effect of pregermination can be measured by evaluating the germinative ability of the stable pregerminated barley kernels. This is done by subjecting the pregerminated and dried kernels to the standard BRF method,[1] where the kernels are regerminated. It consists of placing 100 kernels in a germinative environment (enclosed Petri dish with filter paper and water) and counting the sprouted kernels every 24 h for three days. The germinate energy is then the fraction of germinated kernels out of the 100 kernels. In our case the germinated kernels are counted for five days, using the same conditions as the BRF method.

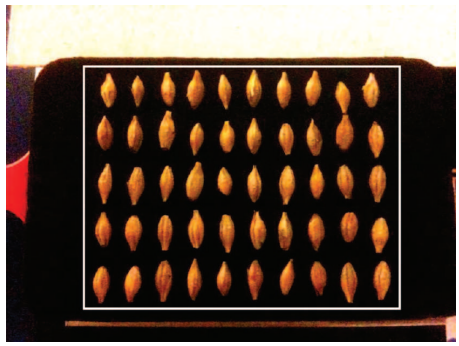


**Figure 3.** Barley kernel sample plate painted with a special NIR black color. A small strip of the NCS-0300 paper is placed on the top as the white reference acquired for each image. Afterward the kernels are cropped, as marked by the white rectangle.

**Image Acquisition and Preprocessing.** *Camera System and Equipment.* A hyperspectral line-scan NIR camera system from Headwall Photonics Inc. (Fitchburg, MA, USA) sensitive in the range 900 to 1700 nm was used to acquire the image of the barley kernels. A dedicated NIR light source illuminates the sample uniformly along the scan line, and an advanced optical system disperses the NIR light onto the InGaAs sensor inside the camera for acquisition. A sledge from MICOS GmbH (Germany) moves the sample past the view slot of the camera allowing it to acquire a hyperspectral image, line by line. The camera has a resolution of 320 spatial pixels and 166 spectral pixels with equally spaced wavelength intervals of 4.8 nm. Further details of the performance of the camera system and the standard operating procedure can be found in the technical report, cf. ref 12.

*Acquisition Procedure.* The procedure for acquiring images can be split into an initialization part to set up the camera system and a subsequent acquisition part. The initialization procedure consists of a series of steps with the purpose to acquire maximum quality images of the barley kernels. Initially the camera system and light source are tuned to maximize dynamic range while limiting specular reflections and saturated pixels. Images of the wavelength standard WSR1920a[13] from Avian Technologies LLC (Sunapee, NH, USA) and a set of vertical lines were acquired for wavelength scale alignment and spatial calibration respectively. The entire process of acquiring images of all the barley kernels once was conducted in one run in a single day.

In order to obtain usable images a maximum of 50 barley kernels were fixated at a time on a sample plate, as shown in Figure 3. The sample plate (courtesy of Skandinavisk Bryggeri Laboratorium, Denmark) is painted with a NIR dark black color (Mankiewicz Nextel-Velvet Coating 811-21 9218 Schwartz) to avoid reflections during acquisitions and for easier background removal during data processing. The acquired images include a white reference [NCS-0300 (Natural Color System, Sweden, Web: http://ncscolour.com) reflecting paper] as the top part in every image to ensure the capture of any potential drift of the light source. An equivalent dark current image is taken at the end of the acquisitions for

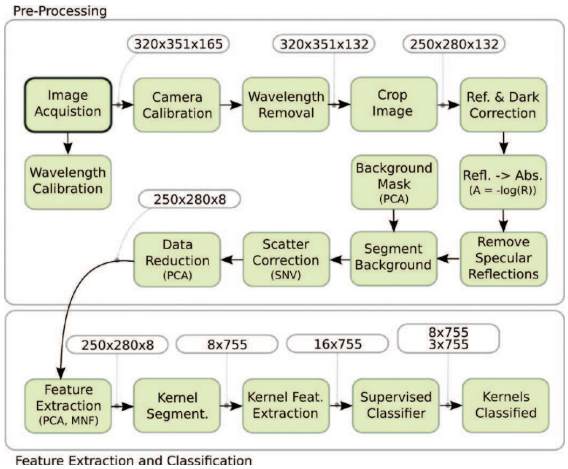



**Figure 4.** Flowchart of the preprocessing, feature extraction, and classification steps starting from the boldface *Image Acquisition* box. The pinned boxes show the size of the image data at the different steps.

each germination time by obstructing the incoming light to the camera with a lens cap.

*Preprocessing Pipeline.* The preprocessing pipeline, as illustrated in Figure 4, is made from a series of steps to prepare the data structure for subsequent analysis. Initially the wavelength scale is recalculated using the wavelength standard WSR1920a[13] to achieve optimal calibration. The camera system suffers from a small misalignment in the optical system, which causes a minor spatial distortion over both the spectral and spatial range. This is corrected using an image of a set of vertical lines and an image of the wavelength standard WSR1920a for calibration. The wavelength range is reduced to 132 bands between 1002 and 1626 nm due to the poor sensitivity of the InGaAs sensor at the spectral extremities. The camera further suffers from a single vertical line of poor sensitivity across the first column of barley kernels, and the affected pixels will represent outliers. These kernels are eliminated from the data set in all images, leaving only 45 kernels per image and a reduced data set of 755 kernels, cf. Table 1. Each acquired image is then represented in a 250 pixels × 280 pixels × 132 bands 3-way tensor after cropping the kernels out (white rectangle in Figure 3). Every image is corrected with the associated white reference and dark current image by $R = (I_{raw} - I_{dark})/(I_{white} - I_{dark})$. The spectral mixing is assumed dominated by subsurface penetration and thus nonlinear multiplicative mixing. To enforce linear mixing the spectra are transformed from reflectance data to absorbance data by $A = -\log_{10}(R)$. Pixels suffering from specular reflections do not hold any spectral information relevant for the analysis. They are identified with a reflection larger than one $R > 1$ or a negative absorbance $A < 0$ and are removed. The background can easily be removed due to the high contrast between the kernels and the NIR absorbing black color by a simple approach. This contrast will exhibit the highest variance and can be captured by a principal component analysis (PCA) decomposition. This is conducted by unfolding each tensor image data into a matrix, where each column represents a pixel, and subjecting the mean centered matrix to PCA. The foreground and background contrast is represented in the first score image and is used to extract a mask from a simple scalar threshold. The acquired images are further corrected for scatter effects as a preprocessing step on the spectral information. Many different scatter correction techniques exist, and we will apply the *standard normal variate* (SNV) correction.[14] Each of the acquired spectra consists of 132 bands, and the relevant information for our analysis is expected to be represented by fewer components. The spectral information is therefore compressed to $D$ dimensions, which is set manually, to suppress the noise allowing for a reduction in the computational load and curse of dimensionality, but at the risk of removing information important for the final analysis. In order to represent all germination times a concatenated data subset is formed from one SNV corrected image from each germination time. This tensor subset is unfolded to a matrix, where each column holds a spectrum. The mean centered data is subjected to PCA, and the entire data set is projected onto the first $D$ loadings.

**Classification Models.** The classification of the individual kernels is based on both spectral and spatial features from the image data, and these features are extracted prior to the classifier (cf. Figure 4).

*Feature Extraction.* Prior to the actual classification, the image data is processed to find features, which exhibit the chemical changes inside the kernels related to the germination process. The feature extraction is applied to a subset of the compressed images represented by one image from each germination time to capture these chemical changes. To this end several feature extraction methods exist, and we will focus on two simple approaches: *PCA*[15] and *minimum noise fraction (MNF)*.[16] PCA finds a linear combination of the spectra, which maximizes the variance in the data and hence does not consider any spatial information. These features have already been extracted in the data compression step and are reused. MNF also finds a linear combination of the spectra by maximizing the signal-to-noise ratio. The signal covariance matrix is based on the spectral information, and the noise covariance matrix is estimated from a neighbor pixel dependency. It thus exploits both spectral and spatial information to estimate components to explain more variance in fewer components than PCA does. The extracted features from any of the decompositions are finally represented as a set of score images.

*Single Kernel Segmentation and Feature Extraction.* The extracted features reveal the chemical changes inside the kernels during germination and need to be quantified prior to the classification. Initially the kernels are segmented out easily since the kernels have the same position

Journal of Agricultural and Food Chemistry ARTICLE

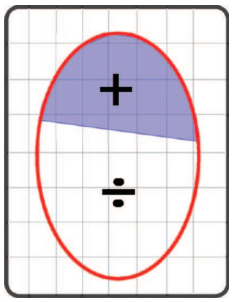


**Figure 5.** The single kernel feature consists of two scalar values for each score image calculated as the sum of the positive and negative intensities respectively divided by the total number of pixels.

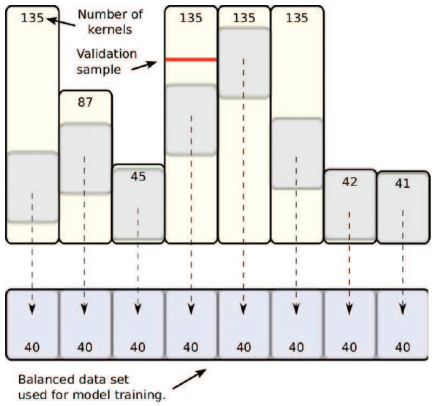



**Figure 6.** Leave-one-out cross validation framework with subsampling to extract training/test sets with equal number of samples from each germination time.

on the sampling plate for all images. In order to quantify the germination progression, a second level of features are extracted from each of the kernels. This feature is based on the spatial information in the extracted score images, as shown in Figure 5, and is constructed to maximize the contrast between the starch break down progression and the background. The feature can be described as the sum of the pixel intensities for each score image for each barley kernel divided by the total number of pixels in the kernel for both positive and negative intensities separately. The advantage of this feature is the insensitivity to the kernel orientation. If we denote the extracted score images $\mathcal{X}^{\text{score}} \in \mathbb{R}^{X \times Y \times D}$, where $X$ and $Y$ are the image dimensions and $D$ is the number of loadings, then the single kernel feature $\mathbf{f}$ can be expressed as

$$\mathbf{f}_{k,n} = \frac{1}{L}[\sum_{x_+} x_{ijk}^{\text{score}} \sum_{x_-} x_{ijk}^{\text{score}}] \quad (1)$$

where $k$ represents the $k$th band, $n$ represents the $n$th kernel, $i,j$ denotes the pixel position and $L$ is the total number of pixels in the kernel. $\mathbf{f}_{k,n}$ is

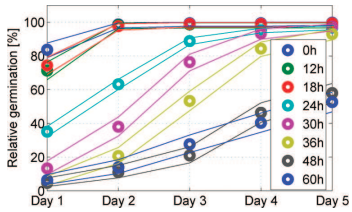



**Figure 7.** Progression of regerminated barley for each day showing the accumulated percentage of germinated kernels for each pregermination duration (dots) with 95% confidence intervals (lines). The progression reveals three clusters of pregermination: normal, delayed and limited.

calculated for each score image $k$ for each kernel $n$ as a 2-dimensional vector, which leads to a complete feature vector $\mathbf{F}_n = [\,\mathbf{f}_{1,n}, \mathbf{f}_{2,n}, ..., \mathbf{f}_{D,n}\,]$ of length $2D$ for each kernel $n$.

*Classification.* We have evaluated different classification models based on different approaches, which includes ordinal classifiers and nonlinear neural networks (cf. ref 11 for supplementary material). The model with the best performance is the maximum likelihood multinomial regression classifier implemented in the L1General toolbox (available for download at http://pages.cs.wisc.edu/gfung/GeneralL1),[17,18] which provides a probability for each barley kernel $n$ for each class $c$ (cf. Figure 4). The classifier can be expressed by the softmax function

$$p(c|\mathbf{F}_n) = \frac{\exp(\mathbf{F}_n \mathbf{w}_c)}{\sum_{\hat{c}=1}^{C} \exp(\mathbf{F}_n \mathbf{w}_{\hat{c}})} \quad (2)$$

where the model weights are denoted by $\mathbf{w}_c$ for each class $c$, and $C$ denotes the total number of classes. The entire data set is applied in a leave-one-out cross validation framework, where the training and test sets are extracted from the entire data set as illustrated in Figure 6 to ensure an equal number samples from each germination time. A different training and test set combination can be extracted repeatedly to achieve an average class probability for each validation sample. In commercial applications bulk results are often desired. Single kernel class probabilities are hence afterward combined as the average class probabilities of adjoining kernels to achieve a more robust probability and provide bulk classification results. A confusion matrix for all $C$ classes is calculated using the class probabilities to evaluate the classification performance. The $C$ classes can be aggregated to a smaller confusion matrix with fewer classes, $\hat{C}$, or the model can be retrained with $\hat{C}$ classes to obtain a more dedicated model with potentially lower classification error. In our case the former approach is used to model the $\hat{C}$ classes.

## RESULTS

This section describes the model and classification results. Supplementary details can be found in the associated technical report.[11]

**Germinative Energy of Barley.** The degrading effect of pregermination is evaluated by subjecting stable pregerminated barley kernels to the standard BRF method,[1] as described previously. The BRF method was repeated several times for each pregermination duration for improved accuracy (cf. Table 1). The resulting progression of the regermination is shown in Figure 7. The regermination results suggest the pregerminated barley can be divided into three groups:

1. Normal regermination. A short pregermination has negligible effect as all kernels regerminate as normal kernels. The period includes 0 to 18 h of pregermination.


dx.doi.org/10.1021/jf202122y |J. Agric. Food Chem. 2011, 59, 11385–11394

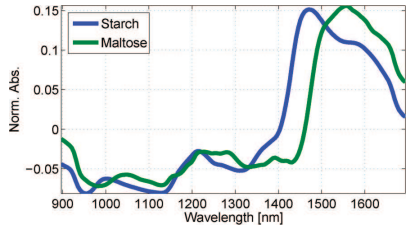



**Figure 8.** NIR spectra of pure constituents of starch and maltose in powder form using our hyperspectral NIR camera system. The two spectra exhibit large differences in the spectral range from 1400 to 1550 nm.

2. Delayed regermination. The kernel will regerminate within five days, but at a slower rate. This period is between 24 to 36 h of pregermination.
3. Limited regermination. The kernels are exhausted from pregermination and not all will regerminate within the five days. This period is greater than 36 h of pregermination.

The division further indicates how the germination starts a major chemical change between 18 and 24 h. Applying the BRF method to pregerminated kernels with visual radicle and epicotyl may fail as the method relies on visual inspection. Long pregermination durations will lead to a visible radicle and epicotyl, and hence identifying if the individual kernel has in fact germinated again becomes difficult. However, this does not change the three pregermination categories identified.

**Preprocessing.** The acquired image data is preprocessed as described in the Preprocessing Pipeline section.

*Background Segmentation and Scatter Correction.* The barley kernels are segmented out by using the simple procedure described earlier. Each hyperspectral image is unfolded into a 132 $\times$ $N$ matrix, where $N = 250 \times 280$ pixels $\approx$ 70000 (excluding specular reflecting pixels). The background mask is extracted from the first loading from the PCA decomposition. After the background removal each kernel is on average represented by approximately 400 pixels. Each spectrum is afterward scatter corrected by applying the SNV transform.[14] A detailed comparison with other scatter correction methods is described in the associated technical report,[11] which shows how the SNV method is superior to the others in this application.

*Data Compression.* The data is compressed as described earlier. The mean centered data subset is represented in a matrix of size $132 \times N_{samples}$, where $N_{samples} = 8$ germination durations $\times$45 kernels per image $\times$ $\sim$400 pixels per kernel $\approx$ 140000 spectra, and subjected to PCA. The number of PCs required to capture the most important information is evaluated from a set of estimated projection loadings. To ensure that the relevant information is retained in the data, while achieving a suitable compression, the entire data is projected onto the first $D = 8$ PC's. These components explain 99% of the variance.

**Data Visualization.** For evaluation of NIR spectra, hyperspectral images of the main constituents during germination, starch and maltose in their pure powder form, were acquired using our camera system, and their mean spectra are illustrated in Figure 8. The starch and maltose spectra reveal similar variation in the lower part of the spectrum, but exhibit large differences at the longer wavelengths. We therefore expect to observe spectral activity during germination in the area between 1400 and 1550 nm. The hyperspectral image data of the barley kernels can be visualized from the score images and the associated spectral loadings from the PCA decomposition conducted in the data compression step. The resulting score images can be used to analyze the internal structure of the barley kernels in general and during germination. Figure 9 illustrates the first five score images and spectral loadings of the second and third PC. The second spectral loading is comparable to the spectra of starch and maltose. Activity at the embryo is clearly captured by the third and fifth score images. The third spectral loading reveals a clear peak around 1430 nm indicating spectral activity in the same area as the spectral difference between starch and maltose. The husk of the barley kernel acts as an attenuating filter and is observed as an intensity difference. However, this is not visible due to the normalizing effect of the SNV scatter correction. The fifth score images for the germination times [0, 12, 24, 36, 48 h] are shown in Figure 10. This PC also captures the germination progression identified as the small area at the embryo increasing over time. The corresponding spectrum includes additional spectral features to increase the contrast of the germination progression compared to PC loading 3. As the third and fifth components designate the starch breakdown into maltose, their associated spectral loadings represent the largest contrast between starch and maltose and not necessarily a pure spectral signature.

A visual subjective evaluation based on several similar images indicates a threshold of approximately 24 h, where the kernels would start the major chemical change. Similarly the same images reveal a very small difference between 48 and 60 h indicating the steady state and hence the second threshold between 36 and 48 h. These properties of germination coincide with the results from the germinative energy analysis using the BRF method.

**Feature Extraction.** The PCA features have already been calculated in the data compression step, and the 8 score images are hence reused. For the MNF decompositions the compressed data subset is represented in a matrix of size $8 \times N_{samples}$, where $N_{samples}$ = 8 germination durations $\times$ 45 kernels per image $\times$ $\sim$400 pixels per kernel $\approx$ 140,000 spectra. Features from the entire data set are then extracted using the estimated loading vectors for each method. The MNF method extracts similar components compared to PCA as shown in Figure 11, where the germination progression is clearly captured in component 1. The germination progression is found in a single component as opposed to PCA leading to a stronger contrast between the germinating embryo and the rest of the kernel. The associated spectrum designates the correlations between the MNF feature loading 1 and the original variables. It is comparable to the fifth PC loading in Figure 10 as it includes similar spectral features to increase the contrast of the germination progression. In general the MNF is considered to perform best in terms of extracting spatial features for the classification. The extracted features used in the following are therefore 8 score images from the MNF decomposition.

**Single Kernel Segmentation and Feature Extraction.** The kernels are segmented out and the single kernel feature to quantify the germination progression $\mathbf{F}_n$ is calculated for each kernel $n$ leading to a $2 \times 8 = 16$-dimensional feature vector per barley kernel. For all barley kernels, the entire feature extracted data set $\mathbf{F}$ becomes a $16 \times 755$ matrix.

Journal of Agricultural and Food Chemistry ARTICLE

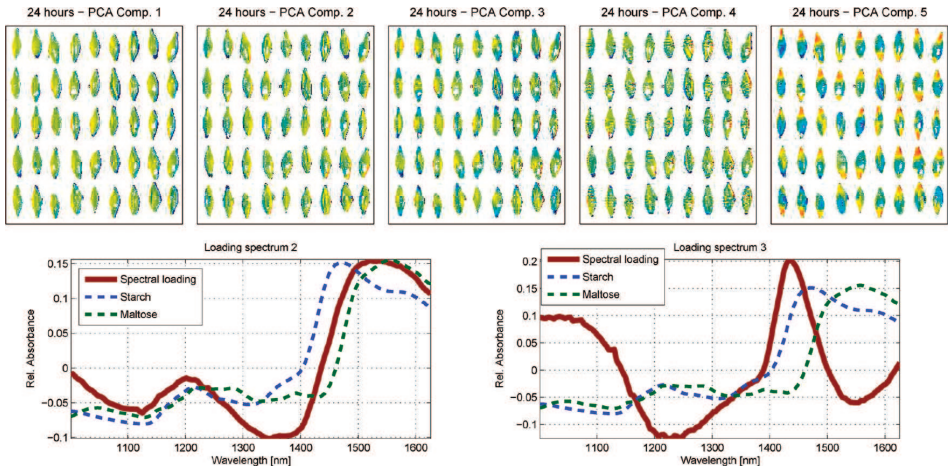



**Figure 9.** PC score images with the corresponding spectral loadings for PC2 and PC3. The second PC loading is comparable to the scaled pure spectra for starch and maltose acquired using the same camera system for comparison (dashed). Both the third and fifth score images have captured activity at the embryo indicating the germination process. The third spectral loading exhibits a distinct peak around 1430 nm and indicates activity in the same area as the spectral signatures of starch and maltose.

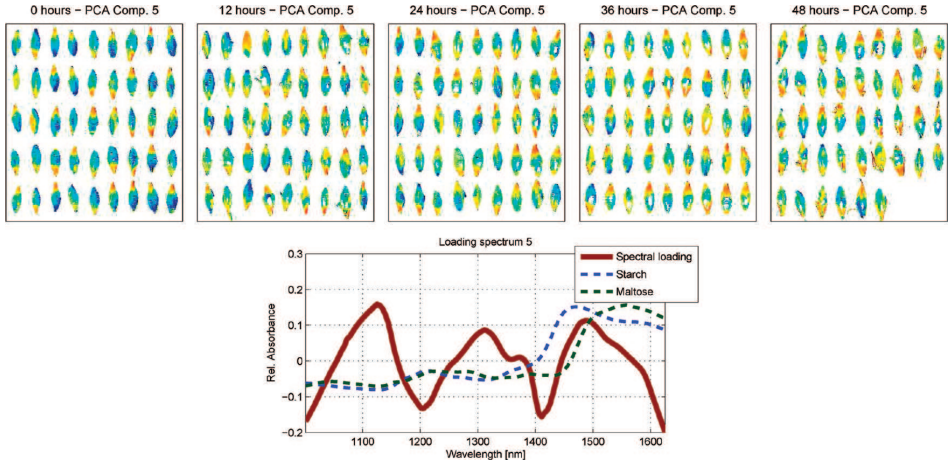



**Figure 10.** PCA score image 5 for the germination durations [0, 12, 24, 36, 48 h] (left to right) and the associated spectral loading. The score images reveal a clear progression in the germination process, and the spectral loading includes additional spectral features to increase the contrast of the germination progression compared to PC loading 3.

**Classifiers.** The classification on a single kernel level will be conducted into the eight classes designating the germination times [0, 12, 18, 24, 30, 36, 48, 60 h] using the MNF feature set.

The leave-one-out approach is applied by extracting a training and test set to ensure equal number of samples from each germination time, i.e. a balanced data set (cf. Figure 6). The germination time with fewest data is 60 h and has only 41 samples (cf. Table 1). Excluding the validation sample the balanced training/test set is hence set to include 40 samples from each germination time, i.e. a total of 320 samples.

Journal of Agricultural and Food Chemistry ARTICLE

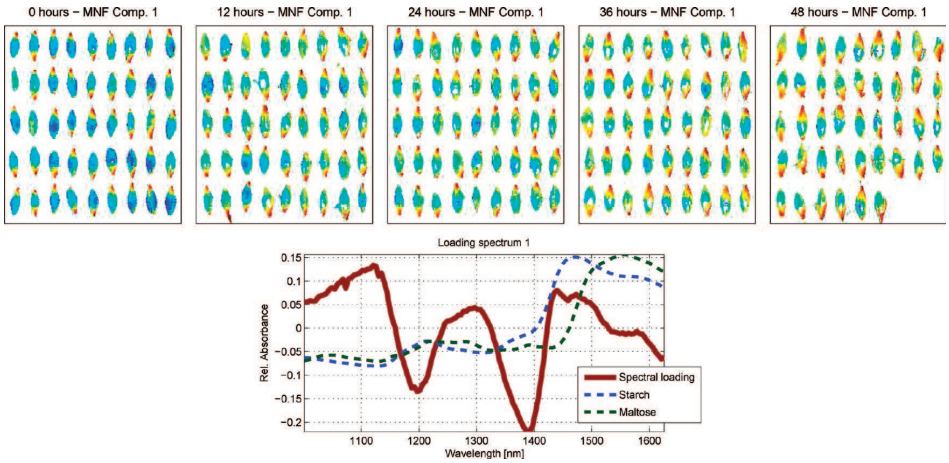



**Figure 11.** The MNF score image 1 on top represents the germination durations [0, 12, 24, 36, 48 h] (left to right) and reveals a stronger intensity of the germinating embryo than PCA does. The spectrum represents the correlations between the MNF feature vector 1 and the original variables. It exhibits spectral activity in the area, where the spectra of starch and maltose differ, and is comparable with PC loading 5 shown in Figure 9.

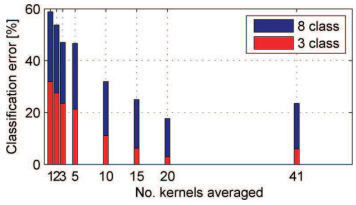



**Figure 12.** Classification errors for different numbers of averaged kernels. The error clearly decreases when averaging more kernels. Averaging kernels leaves fewer samples to analyze and hence higher sensitivity to our results. When averaging 41 kernels, the error is based on only 17 samples.

A random training/test set is extracted 3000 times for each validation sample, a corresponding 8-class classifier model is inferred for each training/test set, and the average class probability for each kernel is used. The classification distribution is calculated in an 8 × 8 confusion matrix to evaluate the classification performance and a corresponding 3 × 3 confusion matrix for the three classes. The validation classification error for all single kernels in all eight classes become 59% ($CI_{95\%}$: 55−62%) and 32% ($CI_{95\%}$: 29−35%) for the three aggregated classes. This can be compared to the random guess error rate of 7/8 = 87.5% and 2/3 = 66.7% respectively.

The class probabilities for adjoining kernels in each class are averaged to estimate bulk classification results for a small range of combinations as shown in Figure 12. The classification error clearly decreases and is a minimum by averaging 20 kernels with an error of 18% ($CI_{95\%}$: 7−35%) and 3% ($CI_{95\%}$: 0−15%) for eight and three classes respectively. Since we only have 755 kernels, averaging them leads to even fewer samples and less accurate results. This is evident for averaging 41 kernels, which only have 17 samples and a higher classification error.

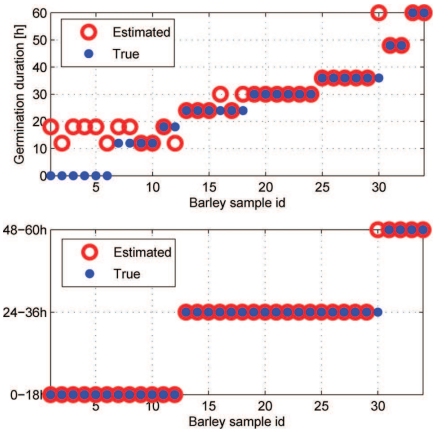



**Figure 13.** Error distribution for averaging 20 kernels for both eight and three classes (top and bottom respectively). The few misclassifications for the eight classes are concentrated around 0−12 h, but stay within the three aggregated classes. The resulting classification error for the three aggregated classes of 3% is due to one misclassification.

The classification distributions for eight and three classes for 20 averaged kernels are illustrated in Figure 13. The misclassifications for the eight classes are concentrated around 0 to 12 h, and are due to the pregermination process having not yet produced macroscopical changes. From this point of view the

Journal of Agricultural and Food Chemistry ARTICLE

pregermination process is with the same stage between 0 and 18 h of pregermination, and causes these misclassifications. The three class validation error of 3% corresponds to one misclassification out of 34 samples and is thus sensitive to the number of samples.

## DISCUSSION

In summary our modeling framework based on hyperspectral NIR imaging technology has proven feasible for describing the degree of pregermination of single barley kernels. The approach of using NIR imaging technology has also proven superior to visually based systems in terms of visualizing the chemical changes inside the kernels during the germination process of barley kernels. The practical germination experiment with eight pregermination levels identified three groups to categorize pregerminated barley: normal, delayed and limited. We presented a simple classifier model to classify the germination durations for single barley kernels. By averaging single kernels to obtain bulk level results and aggregating to three classes we achieved a classification error of 3% ($CI_{95\%}$: 0−15%). Our results are only based on 755 kernels and can be revalidated using a larger data set.

Part of the framework development included evaluation of different approaches, i.e. compression rate and feature extraction method. These choices were made manually and should instead be taken by optimizing the complete classifier framework for a more comprehensive approach. In real-world applications the distribution of single kernels into the three classes is of interest. This gives a more nuanced profile of the pregermination of the kernels than their bulk average. In such applications the classification is performed on a single kernel level and the error of 32% is thus applicable. Our results suffers from having relatively few kernels in the analysis and can be made more accurate and reliable in a future experiment with a larger sample. Our model only assigns pregermination as the cause for a single kernel's lack of germination and is unable to identify dormancy, kernel damage etc. The focus of future work can therefore be to attemp to identify these other properties based on dedicated barley kernel experiments. This can be done in conjuction with our model to build an even more comprehensive framework. The measured water levels of the stored kernels were equal and can be considered a confounder. The impact in the prediction performance should be investigated empirically with other water concentrations. A more robust and difficult approach is to have different water levels for all single kernels to ensure that the model becomes independent of the water levels.

Using hyperspectral images provides the opportunity to achieve a spectrum with an abundance of bands for each pixel. This approach is suitable for research as in this case, but may not be optimal for commercial interests. A natural extension to this work would include the identification of the most important wavelengths to simplify the image technology. If similar results can be achieved with only a few wavelengths, then cost-effective multispectral image equipment can be used instead. Our framework is applicable to other germinating grain types and can possibly be used for other applications such as kernel fungal infections or perhaps detection of dormancy. It is currently based on a single variety of barley and can also be validated against other varieties for possible reinference of model parameters.

Compared to the standard BRF procedure to analyze barley kernels for pregermination taking several days, our proposed automated method is faster and does not include any subjective evaluation. A portion of kernels can be classified within a few minutes. With the system used in this research, image acquistion takes approximately 2 min per sample plate of 50 kernels and the export and classification of data is conducted within tens of seconds. In total the kernels can be analyzed within a few minutes.

The entire analysis is implemented and conducted in MATLAB with an associated toolbox available for download.[19]

## AUTHOR INFORMATION

**Corresponding Author**

*E-mail: ma@imm.dtu.dk, moa@foss.dk pwh@foss.dk; bee@sejet.com; jl@imm.dtu.dk; rl@imm.dtu.dk.

## ACKNOWLEDGMENT

The authors thank Sejet Planteforædling I/S for their assistance in the barley pregermination experiment, Skandinavisk Bryggeri Laboratorium, Denmark, for the sample plates and finally Bo Büchmann (FOSS Analytical A/S) for setting the entire cooperation up and the early review.

## REFERENCES

(1) Germinative Energy of Barley: BRF Method. *Analytica-EBC*; 1997, *Section 3*, Method 3.6.2.

(2) Germinative Capacity of Barley: Rapid Staining Method. *Analytica-EBC*; 2005, *Section 3*, Method 3.5.1.

(3) Pre-Germinated Grains in Barley: Flourescein Dibutyrate Method. *Analytica-EBC*; 2006, *Section 3*, Method 3.8.1.

(4) Bason, M. L.; Ronalds, J. A.; Wrigley, C. W.; Hubbard, L. J. Testing for sprout damage in malting barley using the rapid visco-analyser. *Cereal Chem.* **1993**, *70* (3), 269–272.

(5) Xing, J.; Symons, S.; Shahin, M.; Hatcher, D. Detection of sprout damage in Canada Western Red Spring wheat with multiple wavebands using visible/near-infrared hyperspectral imaging. *Biosyst. Eng.* **2010**, *106* (2), 188–194.

(6) Takeuchi, R.; Kojima, H.; Toyoda, K.; Omoto, H.; Morimoto, S. Morphological measurement of barley seed by computer vision system measurement of changes in seed shape during germination. *Sci. Rep. Fac. Agric. Kobe Univ.* **1992**, *20* (1), 115–121.

(7) Xing, J.; Hung, P. V.; Symons, S.; Shahin, M.; Hatcher, D. Using a Short Wavelength Infrared (SWIR) hyperspectral imaging system to predict alpha amylase activity in individual Canadian western wheat kernels. *Sens. Instrum. Food Qual. Saf.* **2009**, *3* (4), 211–218.

(8) Munck, L.; Moller, B. A new germinative classification model of barley for prediction of malt quality amplified by a near infrared transmission spectroscopy calibration for vigour "on line" both implemented by multivariate data analysis. *J. Inst. Brew.* **2004**, *110* (1), 3–17.

(9) Engelbrecht, P.; Manley, M.; Williams, P. J.; Toit, G. D.; Geladi, P. Pre-germination detected in whole cereal grains using near infrared hyperspectral imaging. *Proc. CST SA—ICC Int. Grains Symp., Qual. Saf. Grain Crops Foods* **2010**, 123–127.

(10) Manley, M.; Engelbrecht, P.; Geladi, P. NIR hyperspectral imaging at Stellenbosch University: detection of pre-germination in whole sorghum grains. *NIR News* **2010**, *21* (7), 11–13.

(11) Arngren, M.; Larsen, J.; Hansen, P. W.; Eriksen, B.; Larsen, R. *Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis, Technical Report*; http://www.imm.dtu.dk/pubdb/p.php?5989; Technical Report, 2010.

(12) Arngren, M. *Hyperspectral NIR Camera System, v1.1*; Technical Report, 2011

(13) Standard Reference Material 1920a (Near Infrared Reflectance Wavelength Standard from 740nm to 2000nm). National Institute of Standards and Technology (NIST).

(14) Barnes, R.; Dhanoa, M.; Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43* (5), 772–777.

(15) Hotelling, H. Relations between two sets of variables. *Biometrika* **1936**, *28* (3/4), 321–377.

(16) Green, A. A.; Berman, M.; Switzer, P.; Craig, M. D. A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal. *IEEE Trans. Geosci. Remote Sens.* **1988**, *26*, 65–74.

(17) Schmidt, M.; Fung, G.; Rosales, R. Fast optimization methods for L1 regularization: A comparative study and two new approaches. *18th Eur. Conf. Mach. Learn., ECML 2007* **2007**, *4701 LNAI*, 286–297.

(18) Schmidt, M.; Fung, G.; Rosales, R. *Optimization Methods for L1-Regularization*; Technical Report, 2009.

(19) Arngren, M.; Larsen, J.; Hansen, P. W.; Eriksen, B.; Larsen, R. *Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis, Matlab Toolbox*; http://www.imm.dtu.dk/pubdb/p.php?5990; Technical Report, 2010.

Appendix F

# Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis, Supplementary material.

Authors: Morten Arngren, Jan Larsen, Per Waaben Hansen, Birger Eriksen, Rasmus Larsen.

Supplementary material for Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis.
http://www2.imm.dtu.dk/pubdb/p.php?5989. Published.

# Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis.

TECHNICAL REPORT v1.0

**Morten Arngren**
Technical University of Denmark, DTU Informatics, Bldg. 321, Richard Petersens Plads, DK-2800 Lyngby
and FOSS Analytical A/S, Slangerupgade 69, DK-3400 Hillerød
ma@imm.dtu.dk,moa@foss.dk, info@arngren.com

**Mikkel N. Schmidt**
Technical University of Denmark, DTU Informatics, Bldg. 321, Richard Petersens Plads, DK-2800 Lyngby
mns@imm.dtu.dk

**Jan Larsen**
Technical University of Denmark, DTU Informatics, Bldg. 321, Richard Petersens Plads, DK-2800 Lyngby
jl@imm.dtu.dk

**Per Waaben Hansen**
FOSS Analytical A/S, Slangerupgade 69, DK-3400 Hillerød
pwh@foss.dkk

**Birger Eriksen**
Sejet Planteforædling, Nørremarksvej 67, DK-8700 Horsens
bee@sejet.com

2

**Version History**

| Version | Date | Change Revision |
|---|---|---|
| 1.0 | Feb. 2011 | Initial issue. |

**Contents**

## 1 Introduction

This technical report acts as an appendix and supplement to the Journal paper entitled 'Analysis of Pre-Germinated Barley using Hyperspectral Image Analysis.' [1]. It is not self-contained as only relevant sections are included in this report.

If you should have any comments or corrections, please forward any inquiry directly to Morten Arngren (ma@imm.dtu.dk or info@arngren.com).

## 2 Scatter Correction

A common effect of image acquisition systems is the light scattering occurring especially for penetrating NIR wavelengths. The scattering effect are mainly be introduced as a subsurface scattering or due to the physical shape of the subject, in this case barley kernels. The scattering is manifested as undesired components in the acquired spectra. This means the scatter correction method acts both as a scatter removal tool in a pre-processing step and as a feature extractor.

This section describes three scatter correction approaches evaluated in detail, ie. *Standard Normal Variate (SNV)*, $0^{th}$ and $1^{st}$ *order polynomial detrending*. Common to these methods is that they operate on the $M$-dimensional observed spectra $\mathbf{S}^{raw} \in \mathcal{R}^{M \times N}$ separately, where $N$ is the number of spectra observed.

### 2.1 Standard Normal Variate

The SNV correction is among the simplest methods, as it both mean centers and normalizes each spectra separately. If we denote a single observed spectrum $\mathbf{s}_i^{raw}$ the SNV correction can be expressed by

$$\mathbf{s}_i^{snv} = \frac{\mathbf{s}_i^{raw} - \mu_i}{\sigma_i}, \tag{1}$$

where $\mu_i$ is the scalar mean value and $\sigma_i$ is the associated standard deviation for the spectra $\mathbf{s}_i^{raw}$.

### 2.2 Detrending

In applying detrending a predefined basis spectrum is subtracted from the observed spectrum $\mathbf{s}_i^{raw}$.

$$\mathbf{s}_i^{detrend} = \mathbf{s}_i^{raw} - \mathbf{s}_i^{basis}, \tag{2}$$

where $\mathbf{s}^{basis}$ denote the basis spectrum, which can be a single background spectrum or consist of a combination of several spectra both dependent and independent of the observed spectrum $\mathbf{s}_i^{raw}$. For certain applications the spectrum of water is known and is removed from all observations if desired. In our case the basis spectrum is made from n'th order polynomials fitted to each observed spectra separately. This means for $0^{\text{th}}$ order detrending the mean constant is subtracted and similarly for $1^{\text{st}}$ order detrending a sloping line is subtracted.

4

2.3 Applied Scatter Correction

The acquired images are subjected to the three scatter correction techniques as a preprocessing step on the spectral information. Figure 1 illustrates a few raw spectra and their scatter corrected equivalents.

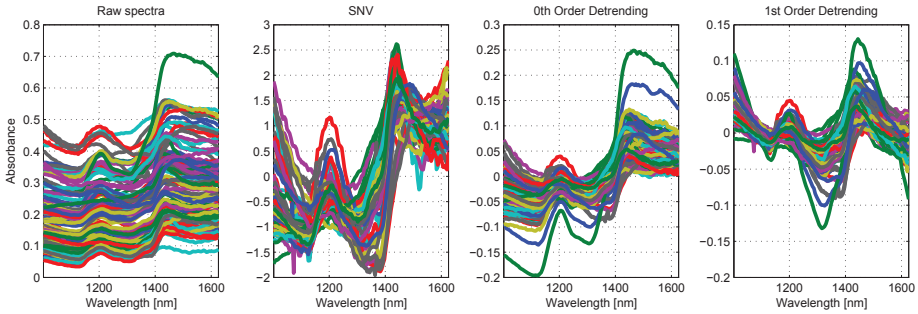



*Fig. 1:* A subset of raw spectra from an acquired image in the left most illustration showing the massive scatter effect. The remaining depictions show the impact of the different scatter corrections.

To compare the scatter correction methods for our application a set of PC score images for each scatter correction method are evaluated in terms of their ability to capture the chemical changes inside the kernels during the germination process. Tensors from one images for each germination time were concatenated and unfolded to a $132 \times N_{samples}$ matrix $\mathbf{X}_{pca}$ such that each mean centered spectrum is represented as a column, where $N_{samples} = 8$ germ. durations $\times$ 45 kernels per image $\times$ 400 avg. pixels per kernel $\approx 140000$ spectra. An unfolded matrix is calculated for each type of scatter corrected data and subjected to PCA providing score images for the evaluation. Figure 2 illustrates the PC score images for the scatter correction methods.

The SNV scatter correction is known to equalize structures in the data due to the scaling and in this case tends to emphasize the edges of the kernels. The only clear captured structure is a strong contrast of the embryo in PC score image 5. The two detrending approaches show similar decompositions. Both has captured the husk of the kernels in the first components as a intensity difference. Activity at the embryo of the kernels is represented in different components. The 0th order pol. detrending extracts this activity in the 3rd and 5th component, whereas the 1st order pol. detrending has captured this feature in the 4th component. Based on the visual evaluation the SNV is used for scatter correction in order to obtain a strong pre-germination feature as separately as possible.

### 3 Data Compression

The acquired spectra consists of 132 bands and the relevant information for our analysis is expected to span a subspace. Regular PCA is hence used to compress the data to reduce the computational load and to suppress the noise, but at the risks of removing information important for the final classification.

A set of projection loadings $\mathbf{U}_{pca}$ are estimated from meancentered spectra from one hyperspectral image from each germination duration becoming a $132 \times \sim 140000$ matrix. The calculated loadings are then afterwards applied to the rest of the data set. Figure 2 (middle) and 3 show the PCA score plots $1-5$ and $7-11$ respectively.


*Fig. 2:* Individual PC score images for each of the scatter corrections: *SNV* (top), *0th order detrending* (middle) and *1st order detrending* (bottom).

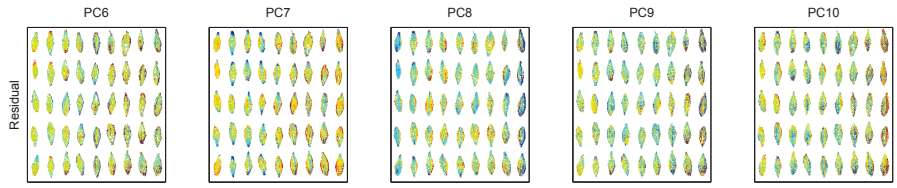



*Fig. 3:* PC score images from 6-10 represent insignificant structure in the image and hence PC's above 8 are disregarded.

The score plots indicate the majority of the structure is covered by the first 8 principal components. To ensure the relevant information is retained in the data, while achieving a suitable compression, the data is projected onto the first 8 PC's. These components further explain 99% of the variance.

6

**4 Feature Extraction**

This section presents the three decompositions, *Principal Components Analysis (PCA)*, *kernel PCA* and *Minimum/Maximum Noise Fraction (MNF)*, used for features extraction to capture the germination progression over time in the acquired hyperspectral images.

### 4.1 Principal Component Analysis

The first and most simple of the three decomposition methods is the principal component analysis (PCA) applied to the spectra. The PCA decomposition maximizes the variance in the spectra and does not consider any spatial information in the images. The extracted PCA features are described in the main article [1].

### 4.2 Kernel Principal Component Analysis (kPCA)

Regular linear PCA is based on inner products to estimate the covariance matrix of the spectral data. The kernel trick framework exploits this fact and can be used to transform the image data into a higher dimensional feature space via a non-linear mapping function $\mathbf{x} \mapsto \mathbf{\Phi}(\mathbf{x})$ applied to the inner products $\mathcal{K} \propto \mathbf{\Phi}\mathbf{\Phi}^\top$ [5]. Applying the kernel trick means the mapping function $\mathbf{\Phi}$ is actually unknown, only the kernel $\mathcal{K}$ is are calculated as the inner products, which in our case is given by the Gaussian kernel

$$k(x_i, x_j) = \exp\big(-\tfrac{1}{2\sigma^2} \parallel x_i - x_j \parallel^2\big), \tag{3}$$

where $\sigma$ is a predefined scaling parameter. Extracting the score images requires processing in the non-linear space due to the mapping function, ie. $\mathbf{\Phi U}$, where $\mathbf{U}$ is the loading vector in feature space. Since this is not directly feasible Larsen *et al.* [5] shows how the projection can be calculated by

$$\mathbf{\Phi U} = \mathbf{K V \Lambda}^{-1/2}, \tag{4}$$

where $\mathbf{V}$ holds the loadings for the kernel covariance matrix $\mathcal{K}$ and $\mathbf{\Lambda}$ represent the associated eigenvalues. The kernel PCA decompositon is memory based, which has the disadvantage of requiring all training data to be available, when projection new data. This means the kernel PCA has a higher computational load when applying to our image data.

The spectral loadings are estimated from a subset of the data consisting of one images from each germination duration, ie. $\mathbf{X}_{train}$ becomes a $8\times \sim 140000$ matrix. Figure 4 depicts the score image showing the clearest progression in the germination process, but it unfortunately includes the other components such as the barley husk and hence does not represent the progression as clean as PCA or MNF score images.

### 4.3 Minimum/Maximum Noise Fraction (MNF)

A different linear decomposition method is the *Minimum/Maximum Noise Fraction (MNF)*. It has received wide attention in the hyperspectral imaging community as it typically represents the structure in the data with fewer components than regular PCA. The MNF approach uses neighbour pixel dependencies and exploits both the spectral and spatial information in the images to estimate more relevant directions for explaining most

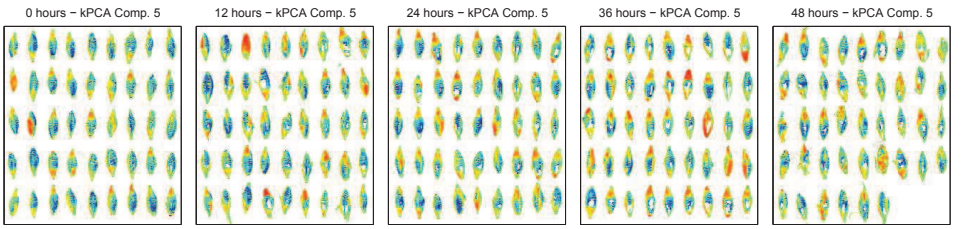



*Fig. 4:* The $5^{\text{th}}$ kPCA score images for the germination durations $[0, 12, 24, 36, 48]$ (left to right). The component does not have the same pure representation as in the PCA or MNF score images.

variance in fewer components. The MNF estimates a set of loading vectors $\mathbf{a}_i$ maximizing the signal-to-noise ratio $\rho$ defined as the Rayleigh quotient expressed by

$$\rho = \frac{\mathbf{a}^\top \Sigma \mathbf{a}}{\mathbf{a}^\top \Sigma_\Delta \mathbf{a}}, \tag{5}$$

where $\Sigma_\Delta$ denote the estimated noise covariance matrix and $\Sigma$ is the sample observation covariance matrix. The noise covariance matrix $\Sigma_\Delta$ is estimated under the assumption that two neighbouring pixels are similar and only differ primarily due to noise. $\Sigma_\Delta$ is then estimated as the neighbour differences for each pixel both vertically and horizontally taking the spatial mask into account and pooled into single covariance matrix [3] [2]. The MNF loading vectors $\mathbf{a}_i$ can then be estimated by simple generalized eigenvalue decomposition [5].

The MNF transform is closely related to the Maximum Autocorrelation Factor (MAF), where the noise covariance matrix is interpreted at the pixel correlation matrix and the estimated loadings point to the maximum autocorrelation.

## 5 Germination Time Classification

The classification of the germination time can be conducted in many different approaches. In the main article the germination times were classified directly by a multinomial classifier. In this section alternative approaches are shortly investigated.

### 5.1 Ordinal Classification

The germination time references exhibits a natural ordering, since class 12h are lower than class 18h etc. Regular classifiers do not utilize this information and risk providing error distributions, where misclassifications are more than one class away from the ground truth. For ordinal data the error distribution should instead only allow misclassifications to neighbour classes.

Regression models incorporates a cost function, where predictions far from the ground truth are penalized hard, e.g. least-squares loss function. The ordinality can therefore be exploited by including a regression term prior to the classification.

8

The ordinal classifier framework includes a prediction step to capture the progression in the references followed by a multinominal regression classifier providing a probability for each class for each barley kernel. The prediction of the germination times are modelled by a regularized non-linear artificial neural network (ANN) [4] [6] [7] [8] [9]. The ANN prediction model is structured as a two layer feed forward network and includes two parameters. Initially the number of hidden units $N_{hu}$ to define the network structure and secondly a regularization parameter $\lambda$ to prevent overfit and force a more linear performance. The `MATLAB` implementation used is the `ANN:DTU Toolbox` from DTU Informatics[1]. The subsequent classification model used is the same linear maximum likelihood multinominal regression classifier as in the main article [1].

The optimal hyperparameters $N_{hu}$ and $\lambda$ for the ANN are estimated in a leave-1-time-out cross validation framework using a single training/test set and applying small range of parameters, $N_{hu} \in [5-18]$ and $\lambda \in [0.5, 1, 2, 5]$. The lowest least-squares prediction error was found for $N_{hu} = 10$ and $\lambda = 1$. Using the optimal hyperparameters, a different training/test set is extracted $100$ times for each validation sample and a corresponding ANN model is inferred for each training/test set. The average germination time predictions illustrated in Figure 6 show how there is a large overlap between the times. During the germination process the starch is broken down at an accelerating rate after approx. $24$ hours, corresponding to the positive slope in Figure 5. This means the kernels within the same germination time can exhibit different germination rates leading to wrong associated time labels. This is seen by the relatively high prediction errors. The last germination time $60$h in Figure 6 is predicted close to $48$h. indicating minor differences in the germination progression.

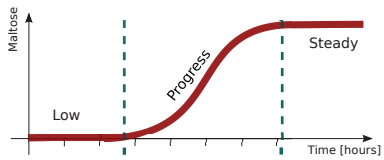



*Fig. 5:* Conceptual illustration of the progression of germinating barley showing the starch break down. The two thresholds indicate the start and stop of the major chemical changes inside the kernel during germination.

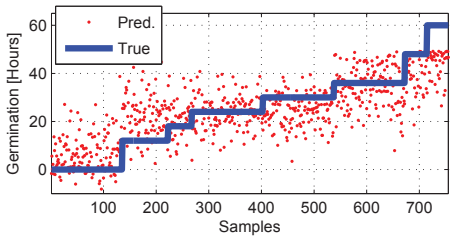



*Fig. 6:* The predicted and true germination time for each of the 755 kernels. The ANN model has captured the overall structure of the germination time despite the large prediction variance. The similar predictions of the last two classes 48h and 60h suggests the germination process begins to level out after 48h of germination.

An associated classifier is inferred for each of the $100$ ANN model using the predicted germination times. The validation classification error for all single kernels in all eight classes become $63\%$ and $33\%$ for the three

[1] Available for download at `http://cogsys.imm.dtu.dk/toolbox/ann/index.html`

aggregated classes. This can be compared to the random guess error rate of $7/8 = 87.5\%$ and $2/3 = 66.7\%$ respectively.

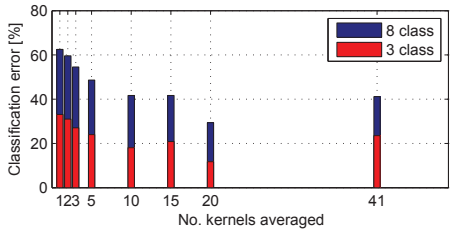



*Fig. 7:* Classification errors for different amount of averaged kernels. The error clearly decreased when averaging more kernels. Averaging kernels leaves fewer samples to analyse and hence higher sensitivity to our results. When averaging 41 kernels the error is based on only 17 samples.

The class probabilities for each kernel can be averaged to estimate bulk classification results for a small range of combinations as shown in Figure 7. The classification error clearly decreases and is minimum by averaging 20 kernels with a error of $29\%$ and $12\%$ for eight and three classes respectively based on $34$ samples. Since we only have $755$ kernels averaging them leads to even fewer samples and less accurate results. This is evident for averaging $41$ kernels, which has only $17$ samples and a higher classification error.

The classification distributions for eight and three classes for $20$ averaged kernels are illustrated in Figure 8. The three class validation error of $12\%$ corresponds to $4$ misclassifications out of $34$ samples and is thus sensitive to the amount of samples.

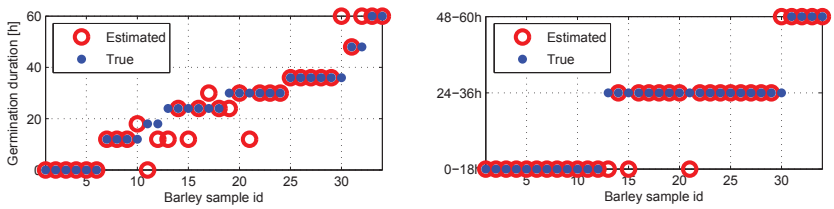



*Fig. 8:* Error distribution for averaging 20 kernels for all eight classes and the three aggregated classes, left and right respectively.

### *5.1.1 Linear Prediction*

The classification of germination times is modelled using an ordinal classifier in order to capture the progression in the references. Our ordinal classifier framework is implemented by conducting the classification on predicted germination times instead of the features directly. This approach ensures the progression is modelled in the prediction step.

In the main article, an Artificial Neural Network (ANN) is used to predict the germination times. A more simple approach is to include a linear model with a regularization term. Using the same notation as in [1], the weights $\mathbf{W}$ of the linear model can be inferred as

$$\mathbf{Y}^{pred} = (\hat{\mathbf{F}}\hat{\mathbf{F}}^{\top} + \alpha\mathbf{I})^{-1}\hat{\mathbf{F}}\hat{\mathbf{Y}}, \tag{6}$$

where $\hat{\mathbf{F}}$ and $\hat{\mathbf{Y}}$ denote the features and references subset used for training respectively, $\mathbf{Y}_{pred}$ are the estimated germination times and $\alpha$ represents the degree of regularization. The regularization term limits any poorly estimated directions in calculating the covariance matrix $\hat{\mathbf{F}}\hat{\mathbf{F}}^{\top}$ due to few samples.

The optimal regularization parameter $\alpha$ is estimated in a leave-1-time-out cross validation framework using $100$ different training/test sets and applying small range of parameters, $\alpha \in [0, 1e9, 1e-8, 1e-7\ 1e-6\ 1e-5\ 1e-4\ 1e-3\ 1e-2\ 1e-1\ 0.2\ 0.5\ 1\ 2]$ as in [1]. The lowest least-squares prediction error was found for $\alpha = 0$, ie. no regularization. Using no regularization, a different training/test set is extracted $3000$ times for each validation sample and a corresponding linear model is inferred for each training/test set.

The average germination time predictions illustrated in 9 show how linear predictions are comparable with the non-linear ANN model. Using the predictions in the same classification framework as before, we achieve a set of higher classification error compared to the ANN model, cf. Figure 10. The error decreases little as the kernels are averaged and does not reveal a strong minimum. The lowest error at $20$ averaged kernels is $59\%$ and $29\%$ for the eight and three classes respectively.

The classification distributions for eight and three classes for $20$ averaged kernels are illustrated in Figure 11. The three class validation error of $29\%$ is by far inferior than the classification error achieved using the ANN predictor.

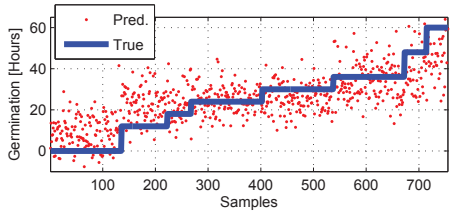



*Fig. 9:* The predicted and true germination time for each of the 755 kernels using the linear predictor. The linear model has captured the overall structure of the germination time despite the large prediction variance and is in general comparable with the ANN model.

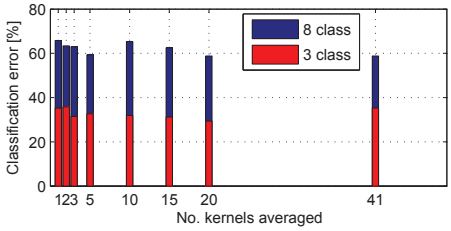



*Fig. 10:* Classification errors for different amount of averaged kernels based on the using the linear prediction model. The error clearly decreased when averaging more kernels. Averaging kernels leaves fewer samples to analyse and hence higher sensitivity to our results. When averaging 41 kernels the error is based on only 17 samples.

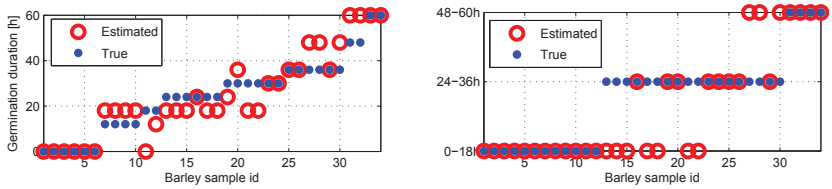



*Fig. 11:* Error distribution for averaging 20 kernels using the linear predictor for all eight classes and the three aggregated classes, left and right respectively.

## 6 Discussion

From a macroscopic viewpoint the $\alpha$-amylase breaks the starch down from the outside and in and hence allows the hyperspectral NIR camera technology to capture these biochemical changes. In contrast if the starch break down would have been initiated from the center of the kernels, it would most likely not have been possible to capture the germination until a later stage in the germination process.

## References

1. Arngren, M., Larsen, J., Hansen, P.W., Eriksen, B., Larsen, R.: Analysis of pre-germinated barley using hyperspectral image analysis. Journal of Agricultural and Food Chemistry (2010)
2. Gordon, C.: A generalization of the maximum noise fraction transform. IEEE Transactions on Geoscience and Remote Sensing **38**(1, part 2), 608–610 (2000)
3. Green, A.A., Berman, M., Switzer, P., Craig, M.D.: A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Transactions on Geoscience and Remote Sensing **26**(1), 65–74 (1988)
4. Larsen, J.: Design of Neural Network Filters. Ph.D. thesis, Electronics Institute, Technical University of Denmark (1993)
5. Larsen, R., Arngren, M., Hansen, P.W., Nielsen, A.A.: Kernel based subspace projection of near infrared hyperspectral images of maize kernels. Lecture Notes in Computer Science **5575/2009**, 560–569 (2009)
6. MacKay, D.J.C.: A practical Bayesian framework for backpropagation networks. Neural Computation **4**, 448–472 (1992)
7. Nielsen, H.: Ucminf - an algorithm for unconstrained, nonlinear optimization. Tech. Rep. IMM-TEC-0019, IMM, Technical University of Denmark (2001)
8. Pedersen, M.: Optimization of Recurrent Neural Networks for Time Series Modeling. Ph.D. thesis, Institute of Mathematical Modeling, Technical University of Denmark (1997)
9. Svarer, C., Hansen, L., Larsen, J.: On Design and Evaluation of Tapped-Delay Neural Network Architectures. In: Proceedings of the 1993 IEEE International Conference on Neural Networks, pp. 46–51. New York, New York (1993)

Appendix G

# Hyperspectral Camera System

Authors: Morten Arngren

Hyperspectral NIR Camera System.
http://www2.imm.dtu.dk/pubdb/p.php?6026, 2011, Published.

# HYPERSPECTRAL NIR CAMERA

Technical Note, ver. 1.2

**by Morten Arngren**

on May, 2011

## Contents

# 1 Hyperspectral Camera

## 1.1 Introduction

This report introduces the Hyperspectral Camera system from Headwall Photonics and describes it in terms of function, performance and potential limitations. From the point of delivery the camera has been subjected to extensive high level testing to evaluate and identify potential performance issue. In addition to presenting the results from the test a standard operation procedure is also given in details.

Throughout the analysis the following nomenclature is used

| | | |
|---|---|---|
| $x$ | : | Scalar |
| $\mathbf{x}$ | : | Column vector |
| $\mathbf{x}_{i,n}$ | : | The n'th datasample at subcoordinate i. |
| $\mathbf{X}$ | : | Matrix |
| $\mathcal{X}$ | : | 3-way tensor |

It should be noted that the light source used to acquire all the hyperspectral images are temporary and not the intended. As all images are compensated for the light source characteristics we do not expect this to have any impact on the evaluation of the camera or any conclusions.

In section 1.2 the camera system is introduced followed by a detailed performance evaluation of the camera in section 1.3. A standard procedure of how to operate the camera describing the steps in detail can be found in section 1.4.

## 1.2 The Camera System

The hyperspectral camera itself is manufactured by the Belgium company XenICs, which also produces the NIR InGaAs sensor used. It is a line-scan camera sensitive from appr. 900-1700nm. meaning it can not acquire an entire image at once. In this camera the vertical axis of the sensor is being reserved for the projected wavelengths and thus only one line can be scanned at the time.

The camera and the setup is shown in Figure 1.1. A dedicated NIR light source illuminates the sample uniformly along the scan line and an advanced optic system developed by Headwall Photonics disperses the NIR light onto the camera sensor for acquisition. This is described in further details in section 1.2.1. A sledge developed by MICOS GmbH moves the sample passed the viewslot of the camera allowing it to acquire a hyperspectral image. This is achieved using a thread driven sledge ensuring high precision movements.

In addition a set of software tools from all manufactures are included to control, calibrate and use the camera system, described below.

- X-control (*XenICs*). Basic camera software to acquire line scans and to calibrate and tune the camera parameters.
- Moco demo (*MICOS GmbH*). Simple software to control the sledge movement.
- Hyperspec Software (*Headwall Photonics*). Image acquisition software based on `LabView` controlling the sledge and line-scan camera to capture images.

The camera system software and hardware includes a series of automated processing steps for the image acquisition, e.g. dead pixel removal, light compensation etc.. In the final set up of the system these automated steps are avoided in order to have maximum control of the entire image processing pipeline. This means additional test must be conducted and a series of `Matlab` function are developed to replace these steps.

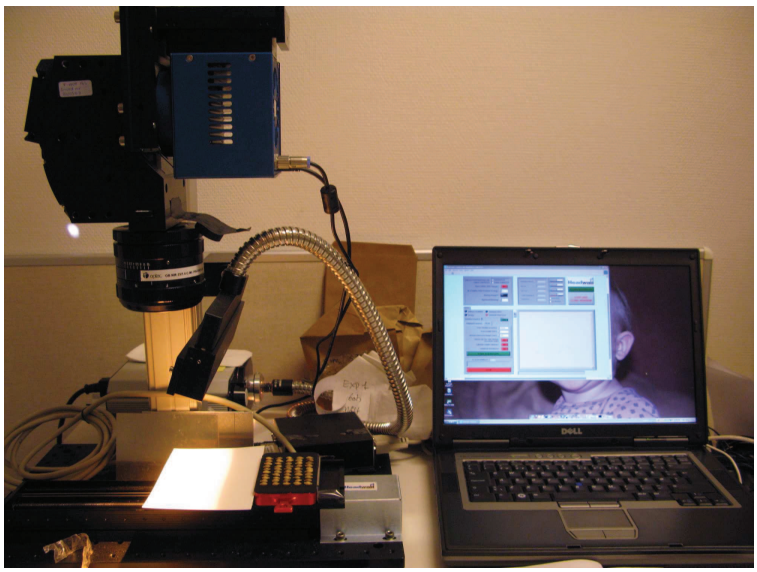


*Figure 1.1:* Mechanical setup of the hyperspectral camera system. (*source: Headwall Photonics Inc.*).

### 1.2.1 Light Dispersion

In order to separate the different wavelengths an optical system based on the Offner principle is used as illustrated in figure 1.2. It consists of a set of mirrors and gratings to guide and distribute the incoming light into a range of wavelengths, which are projected onto the InGaAs sensor.

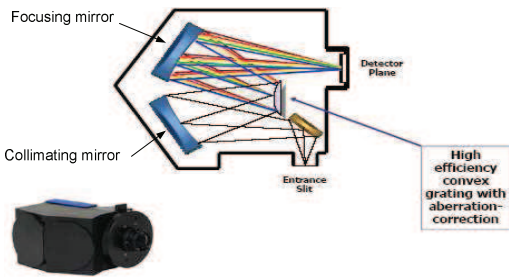



*Figure 1.2:* The Offner principle used to disperse the NIR light and project it onto the InGaAs sensor. (*source: Headwall Photonics Inc.*).

The sensor has a resolution of $320$ spatial pixels and $256$ spectral pixel, i.e. a physical resolution of $320 \times 256$ pixels. Due to the Offner dispersion principle (the convex grating) not all the light are in focus over the entire dispersed range. This means if the light would be dispersed over the whole $256$ pixel wide sensor the wavelengths at the periphery would be out of focus. In order to avoid this the light is only projected onto $166$ pixels instead and the top $90$ pixels are disregarded. This means it essentially becomes a trade-off between spatial sampling resolution and the focus quality of the image.

## 1.3 Camera System Performance

In this section the camera system is subjected to a series of test in order to evaluate the performance in terms of image quality. This analysis forms the basis for the operational procedure described in section 1.4.

### 1.3.1 Camera Calibration

A vital part of the camera system is the possibility to tune or calibrate the camera itself. A small set of parameters regarding the actual acquisition or sampling of the sensor pixels can be adjusted. This allows the user to tune the camera for optimal performance depending on the sample, light conditions, lens, distance to object etc. This however also means the following performance evaluation of the camera is highly dependent on the calibration set used. Prior to any image acquisition for the performance test the camera has been calibrated using the procedure described in section 1.4.3.

### 1.3.2 Sensitivity

The InGaAs sensor is specified to have a sensitivity range of appr. $900-1700$nm. Unfortunately this specification is not based on a lower $\pm 3$db level limit (as for e.g. loudspeakers). Instead the smallest detectable highest and lowest wavelengths are used to designate the active range, as shown in Figure 1.3.

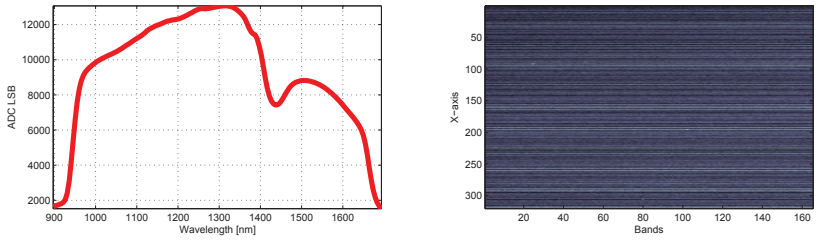



*Figure 1.3:* Average spectra for white reference (NCS-0300) and dark background current image.

The left illustration shows how the sensitivity of the sensor in the periphery is extremely low and with corresponding low Signal-to-Noise ratio (SnR). This is particularly evident from the $900-950$nm range, where almost no light can be detected and the sensor has very poor response. This means the wavelength ranges $900-950$nm and $1650-1700$nm will typically be removed in a pre-processing step and excluded in the further analysis in order to avoid these high absorbances to dominate any signal processing analysis as outliers. The corresponding dark current image shown right in Figure 1.3 clearly reveals the horizontal lines present on the sensor.

### 1.3.3 Dispersion Higher Order Components

In a optical system used in the camera system the incoming NIR light is divided into infinitely many subcomponents denoted by their order as also illustrated in Figure 1.5. The sensor of the camera is then placed and fixed to capture the 1st order component. The dispersion components are however not discrete and will overlap. This means higher order attenuated component are also projected onto the sensor leading to spurious wavelength components.

In order to evaluate the influence of these higher order components and thus the robustness of the system a band blocking filter has been inserted in front of the lens blocking most of the NIR light (the filter characteristics shown in Figure 1.4).

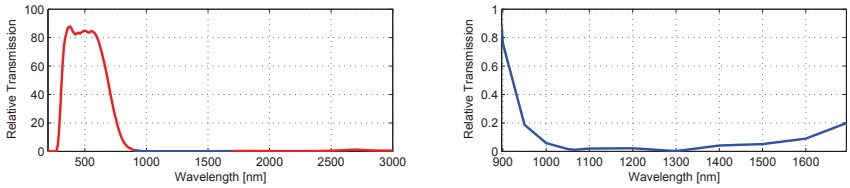



*Figure 1.4:* Transfer function of bandblock filter (zoomed in right figure).

Acquiring a white reference hyperspectral (NCS-0300) with a subsequent dark current correction the average spectra is shown in figure 1.6.

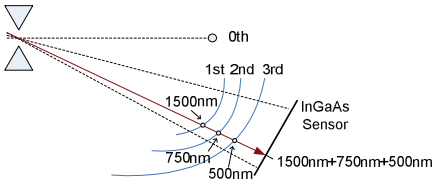



*Figure 1.5:* Principle of dispersion showing sampling of higher order components.

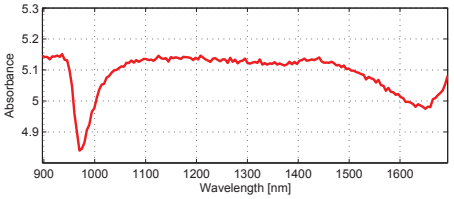



*Figure 1.6:* Camera response to bandblock filter (dark current corrected).

The spectra clearly shows how almost no higher order components has been captured and proves how the system is very robust towards higher order dispersion components. This further omits the need to include a set of filters to remove such higher order components in the system.

### 1.3.4 Noise levels

The sampling resolution of the camera is 14bits defined by the Analog-to-Digital Converters (ADC), i.e. the camera output are integer values ranging from $0-(2^{14}-1)$. As always in real world application the sensor of the camera generates background noise, which may decrease the effective amount of bit to represent the pixel intensity.

The background noise generated is caused by the sensor and the sampling system (pre-amplifiers, ADC's etc.). A common approach to reduce the sensor noise is to cool the sensor prior to image acquisition. The camera is capable of cooling the sensor gradually to lower then $260°$K ($-13°$C) within less than a minute.

The noise level of the sensor can be estimated from a dark current hyperspectral image $\mathcal{X}_{dark} = \{x_{ijk}\}_{ijk=1}^{IJK}$ acquired by obstructing the incoming light to the lens ensuring complete darkness. The average pixel noise $\sigma_{noise}$ is then estimated by

$$\sigma_{noise} = \sqrt{\sum_{ijk}(x_{ijk}-\mu_{ij})^2} \quad \Rightarrow \quad \sigma_{noise} = 11.4\text{lsb} \quad \vee \quad \sigma_{\text{noise}} = \log_2(11.4\text{lsb}) = 3.51\text{bits}, \qquad (1.1)$$

where $\mu_{ij}$ denote the index $i, j$ of the average sensor frame, i.e. the average value for each pixel is subtracted and the standard deviation is estimated as the noise level. This shows the background noise contaminates the lower $3.51$bits. Based on the noise level the corresponding Signal-to-Noise Ratio (SnR) can be estimated as

$$\mathrm{SnR} = \left(1.763 + 6.02 \cdot (14 - \log_2(\sigma_{\mathrm{noise}})\right) \quad \Rightarrow \quad \mathrm{SnR} = 64.9\mathrm{dB} \tag{1.2}$$

Deactivating the cooling and letting the sensor stabilize above the room temperature to appr. $305°$K ($32°$C) and re-estimating the noise levels increases the noise level slightly to appr. $12.6$ LSB decreasing the SnR to $64.1$dB. Hence the cooling should be active for all image acquisitions.

The noise figures estimated are highly dependent on the calibration set used by the camera. A more attractive calibration set can easily be used to reduce the noise even further, but at the expense of failing to capture the dark current. This will leave dark current artefacts in the corrected image and thus reducing the image quality. In addition the calibration is usually set a bit below the max. value in order to avoid saturation leading to a slight noise increase.

### 1.3.5 Oversampling

As part of the camera acquisition parameters the step length or distance between line scans $\Delta m$ can be specified. Normally it is preferred to have square pixels, i.e. same length and width and hence this step length should be specified to obey this (refer to section 1.3.7). In terms of improving the SnR several line scans can be averaged to one by reducing the step length equivalently at the cost of longer acquisition time, e.g. setting $\Delta m \leftarrow \frac{1}{5}\Delta m$ improves the SnR with a factor of $5$ however with the cost of $5$ times the sampling time.

### 1.3.6 Illuminator

A dedicated light source box is part in the camera system to illuminate the sample in the NIR range from 900-1700nm. The illumination system has an associated fiber optic cable and line diffusor ensuring an almost uniform light distribution over the sample. The transfer function of the fiber optic cable and associated line diffusor can be measured by evaluating the difference spectrum between the fiber optical cable being used and bypassed, i.e. direct light source.

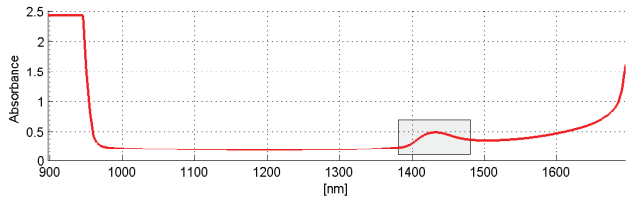



*Figure 1.7:* Optical fiber characteristics.

From figure 1.7 it is clear to see how the absorbance transfer function increases slightly at around 1425nm due to the fiber optic cable. In addition the low sensitivity of the InGaAs sensor at 900-950nm is also very clear to identify.

In such a camera system it is important that the light source is stable during acquisition. If the light source is unstable or drifts over short time then the subsequent reference compensation might induce

fake reflections. It is therefore important to measure the reference reflecting image as part of every image acquisition session.

The light on the sample can be unstable for several reasons. Initially the background light from the room or sunlight can interfere directly. In our case the camera is located in a basement room with no windows and are thus not subjected to sunlight. In order to evaluate the influence from the background room light figure 1.8 illustrates the different spectra of a white reflecting surface (NCS-0300) with and without the room lights on.

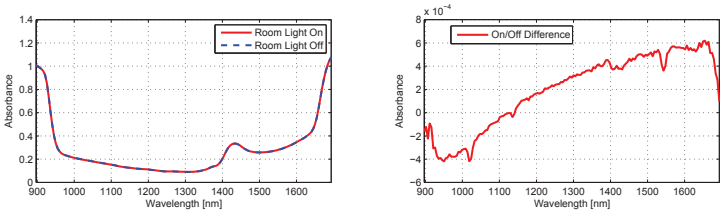



*Figure 1.8:* Absorbance of a white reflecting surface with and without room lights on.

The figure clearly shows how the room light fortunately has negligible influence on the measurements. This means the room light can be left on without degrading image quality.

Secondly the power supply for the light bulb can have poor performance in terms of stability and noise. Any variation here will affect the light emission of the bulb and also the sample illumination. Finally the light bulb itself can drift over time and eventually burn out for replacement. This again emphasizes why a a reference measurement is preferred as part of any real sample acquisition.

### 1.3.7 Acquisition Parameters

As a parameter for the line scan camera setup the distance between line scans must be specified. This parameter $\Delta m$ is dependent on different variables such as the width of the slit $\Delta S$, the distance to the object $\Delta O$ and the focal length $FL$ and can be expressed by

$$\Delta m = \Delta S(\tfrac{\Delta O}{FL} - 1) \tag{1.3}$$

In our setup the following parameters were used: $\Delta S = 25\mu m$, $\Delta O = 250mm$ and $FL = 25mm$. Hence the distance between line scans must be $\Delta m = 225\mu m$. As both the focal length $FL$ and slit width $\Delta S$ are constant the distance between line scans is only dependent on the distance to object parameter and this linear dependency can be illustrated in figure 1.9.

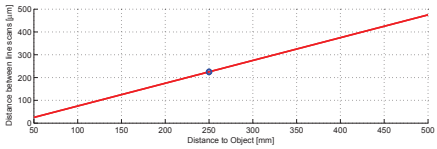



*Figure 1.9:* The distance between line scans depending on the distance to the object.

### 1.3.8 Sensitivity to environment

The different parts of the camera system can fade in their performance over time, temperature or the environment in general. Initially the stability of the light source is vital during image acquisition. If the light source should render unstable due to the bulb itself or the associated power supply, then any image acquired will suffer from undesired artefacts. The light bulb may in addition require a short time to stabilize before image acquisition and may also fade over a longer period of time before replacement. The cooling of the sensor also requires a short time to stabilize. This is however measured live by the acquisition program and can be monitored easily. In order to limit the influence from external variables the camera system must be allowed to stabilize prior image acquisition.

It is therefore important to conduct a acquisition procedure, where the sensitivity towards these external factors are minimized. The best approach is simply to acquire white reference and dark current as often as possible to capture any drift.

### 1.3.9 Dark and Bright reference images

In order to extract the actual NIR response of the samples the influence from both the reference image $\mathbf{X}_{ref}$ and dark current image $\mathbf{X}_{dark}$ are removed. The `Hyperspec` camera software has built-in functionality for this compensation, but in order to have transparency in the entire processing pipeline this step is conducted in `Matlab`.

This means real hyperspectral images for both the reference and dark current images are acquired and the average image frame corresponding to the camera image sensor is used. This also ensures different light illuminations across the line scan will be compensated. Using these two hyperspectral images a single compensated line scan image can be calculated from the measured image $\mathbf{X}_{meas}$ by

$$\mathbf{X}_{ijk}^{(comp)} = \frac{\mathbf{X}_{ijk}^{(meas)} - \mathbf{X}_{jk}^{(dark)}}{\mathbf{X}_{jk}^{(ref)} - \mathbf{X}_{jk}^{(dark)}} \qquad \forall\, i \tag{1.4}$$

where $i, j, k$ denote the indices of the hyperspectral data cube. The dedicated `Matlab` function `prepareCube` can perform the compensation when provided with the hyperspectral reference and dark current images. In order to illustrate figure 1.10 shows typical average reference and dark current line scan image frames.

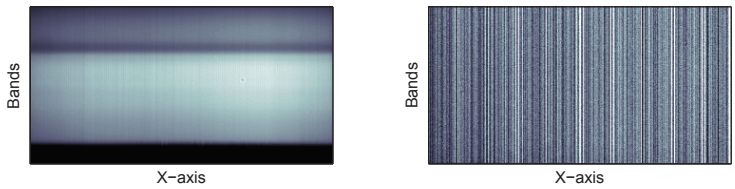



*Figure 1.10:* The average line scan image of a white reflecting surface (*left*) and the dark current image (*right*).

In order to ensure invariance to drift of the illuminator system it is encouraged to capture a white reference for each acquisition and a dark current image for every 5th acquisition as described in the SoP in section 1.4.

### 1.3.10 Spectral Performance

The spectral resolution of the camera is solely determined by the mechanical quality of the optical system. In order to verify the spectral performance a set of simple test can be conducted as described below.

#### Spectral Bandwidth

The spectral bandwidth of the camera is determined by the optical dispersion system and can be estimated from the SRM1920a wavelength standard. By evaluating the 2nd order derivative of the SRM1920a spectrum the bandwidth is determined as the distance between two peaks representing a narrow peak in the original spectrum. This means the peak used must be as narrow as possible in order not to mask the camera's bandwidth. In our case wavelength $f = 1323$nm is relatively narrow and is used for the bandwidth estimation.

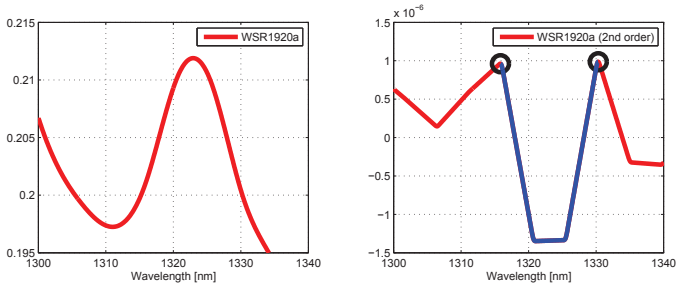



*Figure 1.11:* The bandwidth of the camera based on the 2nd order derivative of the SRM1920a standard.

Figure 1.11 shows the a zoom of the small peak at $f = 1323$nm and the corresponding 2nd order derivative. From this figure the bandwidth is estimated to $\Delta BW_{meas} = 1330\text{nm} - 1316\text{nm} = 14\text{nm}$. This relatively wide bandwidth is slightly larger than the bandwidth of $\Delta BW_{spec} = 9.5$nm specified by the vendors of the camera system.

In this context the large step between each spectral sample of appr. $\Delta f = 4.8$nm must also be considered. This means the measured bandwidth is quantified by the spectral sampling and thus artificially higher.

#### SMR1920a Wavelength Standard

Initially the spectral response of the camera can be compared to the SRM1920a wavelength standard [1] to identify any misalignments of the spectral scale. The acquired average spectra of a SRM1920a sample is shown as the red curve in figure 1.12.

In order to adjust the wavelength scale a simple 1st order function is used to recalculate the wavelength scale expressed by

$$wv_i^* = \alpha \cdot i + \beta \tag{1.5}$$

where $wv_i^*$ denote the new wavelength. The function parameters $\alpha$ and $\beta$ can be estimated using the set of reference points listed in table 1.1 and shown as black circles in the figure.

From these reference points the parameters $\alpha$ and $\beta$ can easily be found (derivations not shown) and the new adjusted wavelength scale calculated. In figure 1.12 the blue curve depicts the corrected SRM1920a spectra with wavelength differences between appr. $4 - 15$nm.

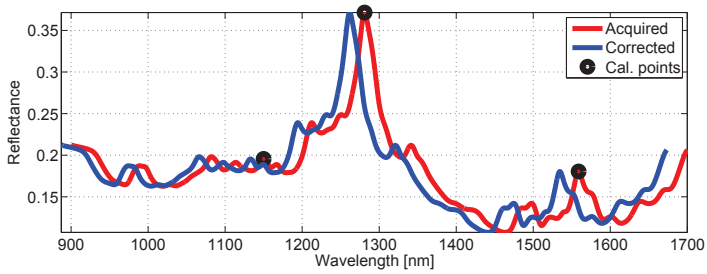



*Figure 1.12:* The wavelength standard used for calibrating the wavelength scale.

| Ref. point | Wavelength |
|---|---|
| 1 | $wv = 1132.9\text{nm}$ |
| 2 | $wv = 1261.8\text{nm}$ |
| 3 | $wv = 1534.6\text{nm}$ |

*Table 1.1:* SRM1920a reference points

The entire calibration functionality is implemented in the `Matlab` function `spectralSRM1920a`, where these reference points are hard coded. The wavelength scale can also be adjusted inside the camera system to ensure a correct wavelength scale being output every time. The approach of post-processing the wavelength scale however ensures an easy and flexible adjustment method.

It is also important to note that this calibration only affect the wavelength scale and not the image data itself, i.e. the quality of the image is not affected if the an image is acquired with a wrong wavelength scale.

Currently three reference points are used as listed in table 1.1. These points are selected based on the bandwidth of the camera system of $10\text{nm}$ according to the SRM1920a standard [1].

### 1.3.11 Intensity Performance

The linearity of the light intensity scale can be evaluated by measuring a range of gray scale images. For this purpose a gray scale set based on the NCS (Natural Color System[1]) color standard is used ranging from NCS-0300 to NCS-9000, where NCS-0300 denotes the most reflecting surface (brightest white) and NCS-9000 denote the darkest surface (deep black).

The different gray scales are measured one by one and compared to a reference measurement using the FOSS XDS instrument, cf. Figure 1.13. The comparison with the XDS reference measurement can also be used to perform a linear 1st order correction of any acquired image, if desired

### 1.3.12 Spatial Performance

The camera suffers from a small misalignment in the optical system causing a minor spatial distortion over both the spectral and spatial range. One of the steps in the pre-processing pipeline is hence to correct for these artefacts.

[1]For a detailed description refer to the NCS website at `http://www.ncscolour.com`

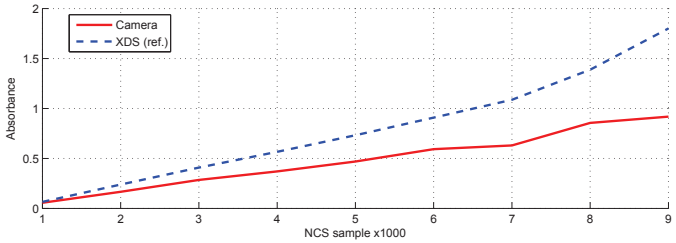



*Figure 1.13:* The light intensity response to different levels of gray.

The distortion can be captured by acquiring an image of vertical lines and the wavelength standard WSR1920a as shown in Figure 1.14 and 1.15. The acquired lines should ideally be strictly vertical, but the average image reveal a large degree of spectral and spatial distortion increasing toward the lower right corner. Similar distortion can be seen in the WSR1920a image at the bottom of Figure 1.15. The distortion is in the range of $2-3$ pixels and is within the optical systems specifications. The optical system does not allow for any mechanical tuning and the correcting must therefore be applied to acquired images as a pre-processing step.

A new set of corrected coordinates are estimated by aligning the positions of the lines in the vertical line image and horizontal lines at the wavelengths $[1132.9, 1261.8, 1534.6]$nm in the WSR1920a image. The correction is afterwards applied to each band frame by linear interpolation.

Correcting an associated dark current image before or after makes no difference due to the linear operation of the interpolation. Figure 1.16 shows a difference image of an acquired barley kernels image before and after the correction. It is clear to see how the correction has compensated most in the lower right corner, as expected.

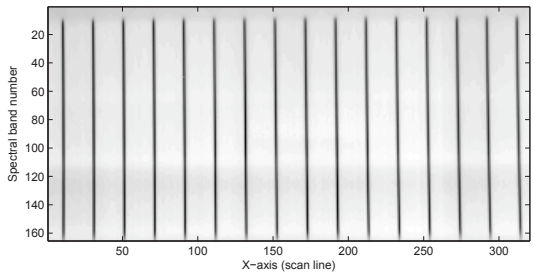



*Figure 1.14:* Averaged image of the vertical lines used for the calibration showing how the lines become more and more skewed to the right of the image.

### 1.3.13 Specular Reflections

Specular reflection occurs when a light source can be seen as a direct reflection on the surface of an object, e.g. sun reflection on a glass ball. Pixels suffering from specular reflection are dominated by the light source and may not contain sufficient information of the reflecting material. In our camera system specular reflections typically occur in the middle of the sample, e.g. grain kernels, due to the position of the light source and camera (cf. Figure 1.1).

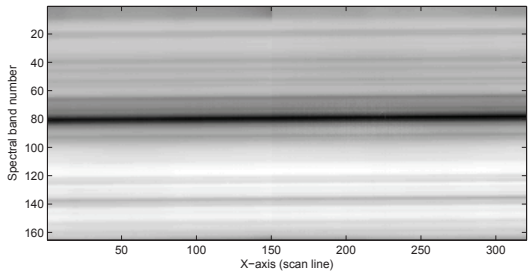



*Figure 1.15:* Averaged image of the wavelength standard WSR1920a used for the calibration showing how the lines are titling a bit to the left.

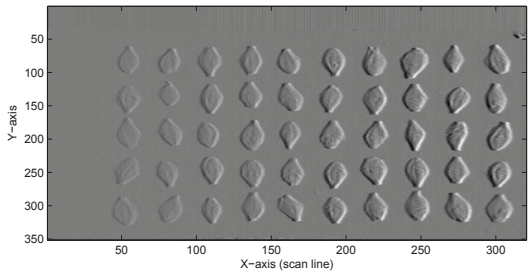



*Figure 1.16:* Averaged difference image of barley kernels between before and after the image correction.

Since the intensity of a specular reflecting pixel is larger than white reference, due to more direct light path, they are easy to identify and mask out subsequently, cf. Figure 1.17.

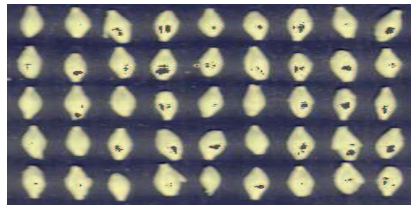


*Figure 1.17:* Barley kernels, where the specular reflecting pixels identified as dark pixel in the middle of the kernels are removed by mask.

### 1.3.14 Dead Pixels

During production of the InGaAs sensor it is common for a small set of the resulting pixels to malfunction in terms of response. This means these bad pixels are either stuck to zero or full response (dark or white) and can easily be identified by capturing a white reference or dark current image.

It is unfortunately not possible to re-activate or fix these bad pixels and since they are stuck at the extrema of the light intensity scale they are considered outliers during a subsequent image analysis processing. The only remedy is to exclude them for the dataset or replace them with the average response of its surrounding neighbours. Dust particles on the lens or in optical system might erroneously be identified as dead pixels. In such a case the optical system will have to be cleaned to remove the dust particles.

The camera has fortunately a built-in compensation for dead pixels by replacing them with the average response of its surrounding neighbours. The mask to identify the dead pixels can be estimated using dedicated camera software. By bypassing the mask dead pixels can be identified by analysing the pixel variances, cf. Figure 1.18.

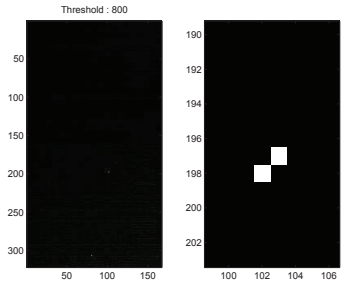



*Figure 1.18:* Sensor frame image identifying a few dead pixels.

For our camera less than $10$ dead pixels are identified corresponding to less than $0.02\%$. Considering the acceptance limit from the camera manufacturer of $0.9\%$, this is by far acceptable. Hyperspectral image acquired does not need to be compensated for dead pixels, as this is already conducted by the camera acquisition software. In addition, the identification of specular pixels can further identify any dead white pixels for out masking.

### 1.3.15 Image Samples

A sample of real-world food samples is shown in Figure 1.19 to illustrate the cameras subjective performance.

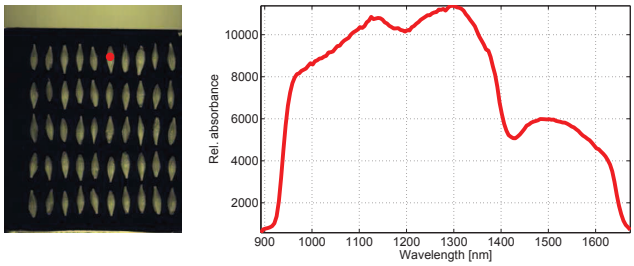



*Figure 1.19:* Example of acquired raw hyperspectral image of barley kernels and associated spectra prior to any pre-processing.

## 1.4 Standard Operational Procedure

This section describes how to acquire hyperspectral images using the camera in a step-by-step procedure. The sequence order to acquire an image is listed below.

- Install camera software (*Only needs to be done once.*).
- Setup the camera mechanics (*Only needs to be done once.*).
- Calibrate camera parameters using `x-control` software (*Only needs to be done once.*).
- Spectral Calibration using the SRM1920a standard.
- Spatial Calibration using squared patterns.
- Compensation for reference and dark current images.
- Acquire Hyperspectral Image.

It is assumes during all steps that the camera is turned on along with the moving sledge and NIR light source. The individual steps are described in details in the following sections.

### 1.4.1 Installing Software

The vendors of the camera and sledge system have their own dedicated software and for the entire system Headwall Photonics has collected additional software components into a single installable program based on `LabView` from National Instruments. To install all software packages use the following procedure.

- Do not connect the camera to the PC yet (this will initiate the search for drivers, which are not installed yet).
- Install XenIC's software by running '*X-Control-Setup-Advanced.exe*'.
- Install MoCo DC software to control the sledge system manually (optional).
- Follow the procedure described in '*CD-1065 Hyperspec A.1.13 Software Manual Rev Y06.pdf*' (installs CD-1079 and CD-1065).
- Go to the '*C:\Program Files\X-Control*' and replace the '*xcamera.dll*' with version 1.2.0.747 (This will fix cooling control of the camera).
- Connect the camera and let Windows install tjhe driver for the camera.
- Start the acquisition software from the '*Hyperspec*' folder in '*C:\Program Files*'.

### 1.4.2 Mechanical Setup

A standard set up of the camera system is assumed during the operational procedure. Deviations are not prohibited, but may give a different quality of the images.

Initially the position of the camera must be as close to the sample without loosing focus in order to maximized the resolution of the samples. The standard position of camera is to be set to $25$cm from lens to object. The camera can be moved by loosening the screws on the side of the vertical bar and adjusting the height of the camera.

Afterwards the focus of the camera lens must be adjusted to the new position. This is achieved by placing a squared paper at the same vertical position as the sample and using the live view in `X-control` software to manually adjust the focus on the lens.

The NIR light source must be turned on during acquisition. Background room light has negligible influence and can be turned on. Sunlight however should be avoided as the influence has not been verified.

### 1.4.3 Calibration of camera parameters

A large set of parameters can be set to tune the performance of the camera. In order to maximize the quality of the images at the expense of acquisition time, the following standard calibration setup is to be used.

- **Integration time** set to max. of $100000\mu s$
- **Cooling Temperature** set to $260°$K
- **Fan Cooling** activated
- **Low Gain Mode** activated

The performance of the camera can be additionally tuned by adjusted the sampling range of the Analog-to-Digital Converters in order to minimize the noise level while maximising the dynamic range. Such a calibration is iterative and can be conducted using dedicated the `X-control` software from the camera vendor XenICs.

In the `X-control` software a histogram of the pixels are shown at the bottom and is very useful for tuning the ADC sampling parameters. In calibrating the camera the darkest and brightest image must be acquired to measure the noise floor and saturation level respectively. It is important to capture the noisefloor in order for subsequential removal. In order to tune use the following iterative steps.

- Set the light source to a high level. This can be adjusted down if necessary during the calibration.
- Place a white reflecting surface under the camera, preferably color sample NCS-0300.
- Adjust $V_{in}$ and $V_{ref}$ to ensure no pixels are saturated.
- Obstruct the incoming light by a dark surface, preferably color sample NCS-9000.
- Adjust $V_{in}$ to ensure the noise floor is within the measurement.
- Repeat the steps until the noise floor is captured while avoiding pixel saturation.

The resulting set of parameters can afterwards be saved as a calibration set and subsequently used in the `Hyperspec` software for image acquisition.

A dedicated calibration menu also exist in the `X-control` software, but it is used for identifying dead pixels on the sensor. The camera has the option to store an image identifying these dead pixels and perform automated averaging during acquisition.

### 1.4.4 Image Acquisition

For the acquisition of a hyperspectral image the `Hyperspec` software is used, all other camera software packages must be shut down. To acquire an image use the following procedure.

- Load the calibration pack.
- Set the '*Distance Between Images*' according to distance to object or oversampling, refer to section 1.3.5 and 1.3.7.
- Deactivate all reference and dark current compensations.
- Place object in sample tray with a NCS-9000 color sample as background or similar.

- The start position of the scanning depends on the position of the sample. This needs to be adjusted by setting the '*Start Position*' and '*Scan Length*' accordingly.
- Activate the '*Create BIL File and Header File from Scan*' to ensure the image is saved.
- Press the '*Start Acquisition*' and enter a filename to save the image to.

The acquired image is now saved and can be viewed by the '*View Image and Spectral Plots*'.

### 1.4.5 Reference and Dark Current Compensation

Any acquired image must be compensated for the reference and dark current images as described in section 1.3.9. To obtain a set of hyperspectral reference and dark current images use the following procedure.

- Use a white reflecting surface as a sample, preferably color sample NCS-0300.
- Acquire the reference image.
- Use a dark surface to obstruct the incoming light to the camera lens, preferably color sample NCS-9000.
- Acquire the dark current image.

## References

[1] Standard reference material 1920a (near infrared reflectance wavelength standard from 740nm to 2000nm). *National Institute of Standards and Technology (NIST)*.

# Bibliography


[1] Germinative energy of barley: Brf method. *Analytica-EBC*, Section 3:Method 3.6.2, 1997.

[2] Germinative capacity of barley: Rapid staining method. *Analytica-EBC*, Section 3:Method 3.5.1, 2005.

[3] Pre-germinated grains in barley: Flourescein dibutyrate method. *Analytica-EBC*, Section 3:Method 3.8.1, 2006.

[4] J. Plaza M. Zortea A. Plaza, G. Martin and S. Sanchez. *Chapter: Recent Developments in Spectral Unmixing and Endmember Extraction, in: Optical Remote Sensing - Advances in Signal Processing and Exploitation Techniques*. Springer, 2010, 2010.

[5] M. Arngren, J. Larsen, P. W. Hansen, B. Eriksen, and R. Larsen. Analysis of pre-germinated barley using hyperspectral image analysis. *Journal of Agricultural and Food Chemistry*, 2010.

[6] M. Arngren, J. Larsen, P. W. Hansen, B. Eriksen, and R. Larsen. Analysis of pre-germinated barley using hyperspectral image analysis, matlab toolbox. url: `http://www.imm.dtu.dk/pubdb/p.php?5990`. Technical report, DTU Informatics, FOSS Analytical A/S, 2010.

[7] M. Arngren, M. Schmidt, and J. Larsen. Bayesian nonnegative matrix factorization with volume prior for unmixing of hyperspectral images. In *Machine Learning for Signal Processing, IEEE Workshop on (MLSP)*, 2009.

[8] M. Arngren, M. N. Schmidt, and J. Larsen. Unmixing of hyperspectral images using bayesian non-negative matrix factorization with volume prior. *Journal of Signal Processing Systems*, 2010.

[9] M. Arngren, M.N. Schmidt, and J. Larsen. Bayesian nonnegative matrix factorization with volume constraints, matlab toolbox. url: `http://www.imm.dtu.dk/pubdb/p.php?5834`. Technical report, DTU Informatics and FOSS Analytical A/S, 2009.

[10] M. Arngren, M.N. Schmidt, and J. Larsen. Nonnegative matrix factorization with volume constraints, matlab toolbox. url: `http://www.imm.dtu.dk/pubdb/p.php?5846`. Technical report, DTU Informatics and FOSS Analytical A/S, 2009.

[11] R.J. Barnes, M.S. Dhanoa, and Susan J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, Vol.43 Issue.5:772–777, 1989.

[12] M. Berman, L. Bishof, and J. Huntington. Algorithms and software for the automated identification of minerals using field spectra or hyperspectral imagery. Technical report, CMIS / CSIRO, 1999.

[13] M. Berman, P.M. Conner, L.B. Whitbourn, D.A. Coward, B.G. Osborne, and M.D. Southan. Classification of sound and stained wheat grains using visible and near infrared hyperspectral image analysis. *Journal of Near Infrared Spectroscopy*, 15(6):351–358, 2007.

[14] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J.F. Huntington. Ice: a statistical approach to identifying endmembers in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 42(10):2085–2095, 2004.

[15] J.M. Bioucas-Dias and J.M.P. Nascimento. Hyperspectral subspace identification. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.46 Issue.8:2435–2445, 2008.

[16] José M. Bioucas-Dias and Antonio Plaza. Hyperspectral unmixing: geometrical, statistical, and sparse regression-based approaches. *Proceedings of SPIE*, Vol.7830 Issue.1, 2010.

[17] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[18] J. W. Boardman, F. A. Kruse, and R. O. Green. Mapping target signatures via partial unmixing of aviris data. *Summaries of JPL Airborne Earth Science Workshop*, 1995.

[19] C. C. Borel and S. A. W. Gerstl. Nonlinear spectral mixing models for vegetative and soil surfaces. *Remote Sensing of Environment*, Vol.47 Issue.3:403, 1994.

[20] Alison J. Burnham, John F. MacGregor, and Roma Viveros. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, Vol.48 Issue.2:167–180, 1999.

[21] Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis*. Chapman and Hall/CRC; 3 edition, 2008.

[22] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, Vol. 46, No. 3.:167–174, 1992.

[23] Tsung-Han Chan, Chong-Yung Chi, Yu-Min Huang, and Wing-Kin Ma. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, Vol.57 Issue.11:4418–4432, 2009.

[24] S. Chandrasekhar. *Radiative Transfer*. Dover Publications Inc., 1960.

[25] Chein-I Chang and Qian Du. Estimation of number of spectrally distinct signal sources in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.42 Issue.3:608–619, 2004.

[26] Xuexia Chen and Lee Vierling. Spectral mixture analyses of hyperspectral data acquired using a tethered balloon. *Remote Sensing of Environment*, Vol.103 Issue.3:338–350, 2006.

[27] Norman B. Colthup, Lawrence H. Daly, and Stephen E. Wiberley. *Introduction to Infrared and Raman Spectroscopy*. Academic Press; 3 edition, 1990.

[28] M.D. Craig. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.32 Issue.3:542–552, 1994.

[29] Stephen R. Delwiche, I-Chang Yang, and Moon S. Kim. Hyperspectral imaging for detection of black tip damage in wheat kernels. *Proceedings of SPIE - The International Society for Optical Engineering*, Vol.7315 Issue.1, 2009.

[30] M. S. Dhanoa, S. J. Lister, R. Sanderson, and R. J. Barnes. The link between multiplicative scatter correction (msc) and standard normal variate (snv) transform of nir spectra. *Journal of Near Infrared Spectroscopy*, Vol. 2:43–47, 1994.

[31] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero. Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery. *IEEE Transactions on Signal Processing*, Vol.57 Issue.11:4355–4368, 2009.

[32] N. Dobigeon, J. Tourneret, and S. Moussaoui. Blind unmixing of linear mixtures using a hierarchical bayesian model. application to spectroscopic signal analysis. *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 79–83, 2007.

[33] Nicolas Dobigeon, Saïd Moussaoui, J.-Y.Jean-Yves Tourneret, and Cédric Carteret. Bayesian separation of spectral sources under non-negativity and full additivity constraints. *Signal Processing*, Vol.89 Issue.12:2657–2669, 2009.

[34] O. Eches, N. Dobigeon, and J.-Y. Tourneret. Estimating the number of endmembers in hyperspectral images using the normal compositional model and a hierarchical bayesian algorithm. *Selected Topics in Signal Processing, IEEE Journal of*, Vol. 4, Issue 3:582, 2010.

[35] D. J. Gardiner. *Practical Raman Spectroscopy*. Springer; 1 edition, 1989.

[36] P. Geladi, D. MacDougall, and H. Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, Vol.39 Issue.3:491–500, 1985.

[37] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[38] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, Vol. 85, Issue 410:398–409, 1990.

[39] Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, Vol.87 Issue.418:523–532, 1992.

[40] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC; 2 edition, 2003.

[41] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, Issue 6:721–741, 1984.

[42] N. Gorretta, J. M. Roger, M. Aubert, V. Bellon-Maurel, F. Campan, and P. Roumet. Determining vitreousness of durum wheat kernels using near infrared hyperspectral imaging. *Journal of Near Infrared Spectroscopy*, 14(4):231–239, 2006.

[43] Andrew A. Green, Mark Berman, Paul Switzer, and Maurice D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74, January 1988.

[44] Robert O. Green, Michael L. Eastwooda, Charles M. Sarturea, Thomas G. Chriena, Mikael Aronssona, Bruce J. Chippendalea, Jessica A. Fausta, Betina E. Pavria, Christopher J. Chovita, Manuel Solisa, Martin R. Olaha, and Orlesa Williams. Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris). *Remote Sensing of Environment*, Vol. 65, Issue 3:227–248, 1998.

[45] B. Hapke. Bidirectional reflectance spectroscopy. i. theory. *Journal of Geophysical Research*, 86(B4):3039–3054, 1981.

[46] Bruce Hapke. *Theory of Reflectance and Emittance Spectroscopy*. Cambridge University Press, 2005.

[47] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, Vol.57 Issue.1:97–109, 1970.

[48] Harold Hotelling. Relations between two sets of variables. *Biometrika*, Vol. 28, No. 3/4:321–377, 1936.

[49] Marian-Daniel Iordache, Antonio Plaza, and Jose Bioucas-Dias. On the use of spectral libraries to perform sparse unmixing of hyperspectral data. *2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2010.

[50] Marian-Daniel Iordache, Antonio Plaza, and Jose Bioucas-Dias. Recent developments in sparse hyperspectral unmixing. pages 1281–1284, 2010.

[51] Tomas Isaksson and Tormod Næs. The effect of multiplicative scatter correction (msc) and linearity improvement in nir spectroscopy. *Applied Spectroscopy*, Vol. 42, Issue 7:1273–1284, 1988.

[52] Steven L. Jacques and Brian W. Pogue. Tutorial on diffuse light transport. *Journal of Biomedical Optics*, Vol. 13(4):041302, 2008.

[53] P.E. Johnson, M.O. Smith, S. Taylor-George, and J.B. Adams. A semiempirical method for analysis of the reflectance spectra of binary mineral mixtures. *Journal of Geophysical Research*, 88(B4):3557–3561, 1983.

[54] N. Keshava. A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal*, 14(1):55–78, 2003.

[55] N. Keshava and J.F. Mustard. Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57, 40 2002.

[56] P. Kubelka and F. Munk. Reflection characteristics of paints. *Zeitschrift fur Technische Physik*, Vol.12:593–601, 1931.

[57] David Landgrebe. Information extraction principles and methods for multispectral and hyperspectral image data. *Information Processing for Remote Sensing*, 1998.

[58] Rasmus Larsen, Morten Arngren, Per Waaben Hansen, and Allan Aasbjerg Nielsen. Kernel based subspace projection of near infrared hyperspectral images of maize kernels. *Lecture Notes in Computer Science*, 5575/2009:560–569, 2009.

[59] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[60] J.B. Lee, A.S. Woodyatt, and M. Berman. Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.28 Issue.3:295–304, 1990.

[61] Jun Li and J.M. Bioucas-Dias. Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data. *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, Vol.3:250–253, 2008.

[62] S. Liangrocapart and M. Petrou. Feasibility study on the use of non-linear spectral unmixing. *Image and Signal Processing for Remote Sensing V*, Vol.3871:159–168, 1999.

[63] R. D. Maesschalck, F. Estienne, J. Verdu´-Andre´s, A. Candolfi, V. Centner, F. Despagne, D. Jouan-Rimbaud, D. L. Walczak, B.; Massart, S. deJong, O. E. deNoord, C. Puel, and B. M. G. Vandeginste. The development of calibration models for spectroscopic data using principal component regression. *Internet Journal of Chemistry 2*, 19, 1999. URL: http://www.ijc.com/articles/1999v2/19/.

[64] S. Mahesh, A. Manickavasagan, D.S. Jayas, J. Paliwal, and N.D.G. White. Feasibility of near-infrared hyperspectral imaging to differentiate canadian wheat classes. *Biosystems Engineering*, 101(1):50–57, 2008.

[65] Harald Martens and Tormod Næs. *Multivariate Calibration.* John Wiley & Sons, 1989.

[66] Yahya M. Masalmah. *Unsupervised Unmixing Of Hyperspectral Imagery Using The Constrained Positive Matrix Factorization.* PhD thesis, Computing And Information Science And Engineering, University Of Puerto Rico, Mayagez Campus, 2007.

[67] Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, Vol. 21:1087–1092, 1953.

[68] L. Miao and H. Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(3):765–777, 2007.

[69] Lidan Miao and Hairong Qi. A constrained non-negative matrix factorization approach to unmix highly mixed hyperspectral data. *2007 IEEE International Conference on Image Processing*, Vol.2:185–188, 2007.

[70] A. Mohammad-Djafari. A bayesian approach to source separation. *AIP Conference Proceedings*, Issue.567:221–244, 2001.

[71] S. Moussaoui, D. Brie, O. Caspary, and A. Mohammad-Djafari. A bayesian method for positive source separation. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.5:V–485, 2004.

[72] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret. Separation of non-negative mixture of non-negative sources using a bayesian approach and mcmc sampling. *IEEE Transactions on Signal Processing*, Vol.54 Issue.11:4133–4145, 2006.

[73] S. Moussaoui, C. Carteret, D. Brie, and A. Mohammad-Djafari. Bayesian analysis of spectral mixture data using markov chain monte carlo methods. *Chemometrics and Intelligent Laboratory Systems*, Vol.81 Issue.2:137–148, 2006.

[74] Michael Myrick, Michael Simcock, Megan Baranowski, Heather Brooke, Stephen Morgan, and Jessica McCutcheon. The kubelka-munk diffuse reflectance formula revisited. *Applied Spectroscopy Reviews*, Vol. 46, Issue 2:140–165, 2011.

[75] M. Mørup and L. K. Hansen. Archetypal analysis for machine learning. In *Machine Learning for Signal Processing (MLSP), IEEE Workshop on*, 2010.

[76] Tormod Naes, Tomas Isaksson, and Bruce Kowalski. Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, Vol. 62 (7):664–673, 1990.

[77] Jos M. P. Nascimento and Jos M. Bioucas-Dias. Nonlinear mixture model for hyperspectral unmixing. *Proceedings of SPIE - The International Society for Optical Engineering*, Vol.7477 Issue.1, 2009.

[78] Jos M. P. Nascimento and Jos M. B. Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.43 Issue.4:898–910, 2005.

[79] José M. P. Nascimento and José M. Bioucas-Dias. Unmixing hyperspectral intimate mixtures. *Proceedings of SPIE–the international society for optical engineering*, Vol. 7830, Issue 1, 2010.

[80] A A Nielsen. Kernel maximum autocorrelation factor and minimum noise fraction transformations. *IEEE Transactions on Image Processing*, Vol. 20, Issue 3:612–624, 2011.

[81] Allan Aasbjerg Nielsen. Orthogonal transformations. Technical report, IMM / DTU, 1999.

[82] Allan Aasbjerg Nielsen, Knut Conradsen, and James J. Simpson. Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, Vol. 64, No. 1:1–19, 1998.

[83] M. F. Ochs, R. S. Stoyanova, F. Arias-Mendoza, and T. R. Brown. A new method for spectral decomposition using a bilinear bayesian approach. *Journal of Magnetic Resonance*, Vol.137 Issue.1:161–176, 1999.

[84] Lucas Parra, Clay Spence, Paul Sajda, Andreas Ziehe, and Klaus-Robert Mller. Unmixing hyperspectral data. *Neural Information Processing Systems*, Vol. 12, 1999.

[85] V. Paul Pauca, J. Piper, and Robert J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications*, 416(1):29–47, 2006.

[86] Lars Petersen, Pentti Minkkinen, and Kim H. Esbensen. Representative sampling for reliable data analysis: Theory of sampling. *Chemometrics and Intelligent Laboratory Systems*, Vol. 77, Issues 1-2:261–277, 2005.

[87] A. Plaza and C.-I. Chang. Impact of initialization on design of endmember extraction algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.44 Issue.11 Part.2:3397–3407, 2006.

[88] A. Plaza, P. Martinez, R. Perez, and J. Plaza. Spatial/spectral endmember extraction by multidimensional morphological operations. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.40 Issue.9:2025–2041, 2002.

[89] Antonio Plaza, Pablo Martnez, Rosa Prez, and Javier Plaza. A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 42(3), 2004.

[90] M.T. Rahman and M.S. Alam. Nonlinear unmixing of hyperspectral data using bdrf and maximum likelihood algorithm. *Proceedings of the SPIE - The International Society for Optical Engineering*, Vol.6566, 2007.

[91] Terrill W. Ray and Bruce C. Murray. Nonlinear spectral mixing in desert vegetation. *Remote Sensing of Environment*, Vol.55 Issue.1:59–64, 1996.

[92] Åsmund Rinnan, Frans van den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, Vol.28 Issue.10:1201–1222, 2009.

[93] D.M. Rogge, B. Rivard, Jinkai Zhang, and Jilu Feng. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.44 Issue.12:3725–3736, 2006.

[94] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, Vol.36 Issue.8:1627–1639, 1964.

[95] Reinhard Schachtner, Gerhard Pppel, Ana Maria Tom, and Elmar W. Lang. Minimum determinant constraint for non-negative matrix factorization. *Lecture Notes in Computer Science*, Vol. 5441/2009:106–113, 2009.

[96] Mikkel N. Schmidt and Hans Laurberg. Nonnegative matrix factorization with gaussian process priors. *Computational Intelligence and Neuroscience*, page Article ID 361705, 2008.

[97] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, Vol. 10, Issue 5:1299–1319, 1998.

[98] J. J. Settle and N. A Drake. Linear mixing and the estimation of ground cover proportions. *International Journal of Remote Sensing*, 14:1159–1177, 1993.

[99] Muhammad A. Shahin and Stephen J. Symons. Detection of hard vitreous and starchy kernels in amber durum wheat samples using hyperspectral imaging (grl number m306). *NIR News*, 19(5):16–18, July/August 2008.

[100] G. Shaw and D. Manolakis. Signal processing for hyperspectral image exploitation. *IEEE Signal Processing Magazine*, 19(1):12–16, 2002.

[101] C.B. Singh, D.S. Jayas, J. Paliwal, and N.D.G. White. Detection of insect-damaged wheat kernels using near-infrared hyperspectral imaging. *Journal of Stored Products Research*, Vol.45 Issue.3:151–158, 2009.

[102] M. O. Smith, P. E. Johnson, and J. B. & Adams. Quantitative determination of mineral types and abundances from reflectance spectra using principal components analysis. *(Lunar and Planetary Institute, NASA, American Geophysical Union, et al., Lunar and Planetary Science Conference, 15th, Houston, TX, Mar. 12-16, 1984) Journal of Geophysical Research, Supplement (ISSN 0148-0227)*, Vol. 90:C797–C804, 1985.

[103] Ben Somers, Kenneth Cools, Stephanie Delalieux, Jan Stuckens, Dimitry Van der Zande, Willem W. Verstraeten, and Pol Coppin. Nonlinear hyperspectral mixture analysis for tree cover estimates in orchards. *Remote Sensing of Environment*, Vol.113 Issue.6:1183–1193, 2009.

[104] Da-Wen Sun, editor. *Infrared Spectroscopy For Food Quality Analysis And Control*. Elsevier, 2009.

[105] Lixin Sun, Ying Zhang, and B. Guindon. Improved iterative error analysis for endmember extraction from hyperspectral imagery. *Proceedings of the SPIE - The International Society for Optical Engineering*, Vol.7086:70860S (10 pp.), 2008.

[106] G. Swayze, R. N. Clark, F. Kruse, S. Sutley, and A. Gallagher. Ground-truthing aviris mineral mapping at cuprite, nevada. *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop.*, Volume 1: AVIRIS Workshop:47–49, 1992.

[107] A. A. Switzer, P. & Green. Min/max autocorrelation factors for multivariate spatial imagery (tech. report 6). Technical report, Stanford University, CA, 1984.

[108] Julius T. Tou and R. C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley; 2nd edition, 1977.

[109] Lihong V. Wang and Hsin i Wu. *Biomedical Optics: Principles and Imaging*. Wiley-Blackwell, 2007.

[110] Michael E. Winter. N-findr: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. *Proceedings of SPIE - The International Society for Optical Engineering*, Vol.3753:266–275, 1999.

[111] L. Zhang, D. Li, Q. Tong, and L. Zheng. Study of the spectral mixture model of soil and vegetation in poyang lake area, china. *International Journal of Remote Sensing*, Vol.19 Issue.11:2077–2084, 1998.